Social Context in Usability Evaluations: Concepts, Processes and Products

Jensen, Janne Jul

Publication date:
2009

Document Version
Accepted author manuscript, peer reviewed version

Social Context in Usability Evaluations: Concepts, Processes and Products

by

Janne Jul Jensen
M.Sc., Aalborg University (2003)

Submitted to the Faculty of Engineering and Science
in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

at

Aalborg University, Denmark

This thesis was defended at the Faculties of Engineering, Science and Medicine, Aalborg University, Denmark on

September 9th, 2009

Opponents:

Professor Ebba Thora Hvannberg
University of Iceland
Department of Computer Science

Associate Professor Panos Markopoulos
Technische Universiteit, Eindhoven
Faculty of Industrial Design

Professor Jan Stage
Aalborg University
Department of Computer Science

Supervisor:

Associate Professor Mikael B. Skov
Aalborg University
Department of Computer Science

Social Context in Usability Evaluations: Concepts, Processes and Products

Janne Jul Jensen

Abstract: This thesis addresses social context of usability evaluations. Context plays an important role in usability evaluations. A major part of the context of a usability evaluation is the people involved. This is also often referred to as the social context of the usability evaluation, and although social context is considered important, only little research has been done to identify how it influences usability evaluations. In this thesis I explore how social context affects the process and product of a usability evaluation and explain the findings in terms of the theory of behaviour settings originating from environmental psychology.

This thesis consists of five published paper contributions and a summary. In the summary I motivate three research questions addressing three aspects of social context. These research questions are answered through a literature review, four laboratory experiments and a field experiment. Findings from these activities are presented in five published paper contributions. I furthermore introduce the theory of behaviour settings as a tool to help characterise the key concepts of social context which, together with an understanding of usability evaluations, provide the framework spanning my research. I then present and discuss the research methods applied in my research, followed by a conclusion on my three research questions including limitations.

The primary results of my research are: 1. Applying the concept of operatives (single leader, multiple leader and joint leader) and non-operatives (members, spectators, neutrals and potentials) from the theory of behaviour settings to usability evaluations generates an understanding and create an awareness of the level of power possessed by each of the participants in the social context. 2. On the operative level, the verbalisation and collaboration of multiple leaders in usability evaluations are affected by acquaintance, and a break down in collaboration or a decrease in verbalisation may cause the test leader to dynamically switch role during the usability evaluation to compensate. However, the influence of non-operatives is subject to some uncertainty. 3. A careful composition of social context can successfully support problem identification. However, problem identification differs between user groups as well as between usability evaluation setups.

Social kontekst i brugbarhedsevalueringer: Begreber, processer og produkter

Janne Jul Jensen

Resumé: Denne afhandling omhandler den sociale kontekst for brugbarheds-evalueringer. Kontekst spiller en vigtig rolle i brugbarhedsevalueringer. En betydningsfuld del af konteksten for en brugbarhedsevaluering er de involverede personer. Disse refereres også ofte til som den sociale kontekst for en brugbarhedsevaluering, og selvom social kontekst anses for vigtig, er der kun i begrænset omfang forsket i hvordan den influerer brugbarhedsevalueringer. I denne afhandlinger undersøger jeg hvorledes social kontekst påvirker processen og produktet af en brugbarheds-evaluering og forklarer resultaterne i termer af teorien om behaviour settings som stammer fra environmental psychology.

Denne afhandling består af fem publicerede artikelbidrag og en sammenfatning. I sammenfatningen motiverer jeg tre forskningsspørgsmål som adresserer tre aspekter af social kontekst. Disse forskningsspørgsmål besvares gennem et litteraturstudie, fire laboratorieeksperimenter og et eksperiment i felten. Resultater fra disse aktiviteter præsenteres i fem publicerede artikelbidrag. Endvidere introducerer jeg teorien om behaviour settings som et værktøj til at karakterisere nøglebegreberne for social kontekst, som, i sammenhæng med forståelsen af brugbarhedsevalueringer, ligger til grund for den struktur der udspænder min forskning. Derefter præsenterer og diskuterer jeg de i forskningen anvendte forskningsmetoder, efterfulgt a en konklusion på mine tre forskningsspørgsmål samt forskningens begrænsninger.

De primære resultater fra min forskning er: 1. Anvendelsen af operatives (single leader, multiple leader and joint leader) og non-operatives (members, spectators, neutrals and potentials) fra teorien om behaviour settings i brugbarhedsevalueringer genererer en forståelse og skaber en opmærksomhed omkring level of power for hver af deltagerne i den sociale kontekst. 2. På operatives niveau påvirkes verbaliseringen og samarbejdet mellem multiple leaders i brugbarhedsevalueringer af det indbyrdes kendskab, og et nedbrud i samarbejdet eller en nedgang i verbaliseringen kan medføre at testlederen dynamisk skifter rolle under brugbarhedsevalueringen for at kompensere for dette. Dog er påvirkningen fra non-operatives genstand for en vis usikkerhed. En velvalgt sammensætning af den sociale kontekst kan succesfuldt understøtte problemidentifikation. Dog varierer problemidentifikationen mellem forskellige brugergrupper såvel som mellem forskellige opsætninger af brugbarhedsevalueringer.

Preface and Acknowledgements

The focus of this thesis is the concepts, processes and products of social context in usability evaluations. The thesis consists of this summary and five individual paper contributions published as follows:

1. Als, B. S., Jensen, J. J. & Skov, M. B. (2005) Exploring Verbalization and Collaboration of Constructive Interaction with Children. *Proceedings of the 10th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'05)*, 443-456, Berlin: Springer-Verlag.

2. Høegh, R. Th. & Jensen, J. J. (2008) A Case Study of Three Software Projects: Can Software Developers Anticipate the Usability Problems in their Software? *Behaviour & Information Technology (BIT)*, 27(4) 307-312, Taylor & Francis Group.

3. Als, B. S., Jensen, J. J. & Skov, M. B. (2009) Composing Children Dyads in Constructive Interaction: A Comparison of Usability Testing Methods for Problem Identification. (Extended version of Als, B. S., Jensen, J. J. & Skov, M. B. (2005) Comparison of Think-Aloud and Constructive Interaction in Usability Testing with Children. *Proceedings of the 4th International Conference for Interaction Design and Children (IDC'05)*, 80-87, New York: ACM Press.)

4. Jensen, J. J. (2007) Evaluating in a Healthcare Setting: A Comparison between Concurrent and Retrospective Verbalisation. *Proceedings of the 12th International Conference on Human-Computer Interaction - Interaction Design and Usability*, 508-516, Berlin: Springer-Verlag.

5. Jensen, J. J. & Skov, M. B. (2009) A Classification of Research Methods and Purposes in Child-Computer Interaction. (Extended version of Jensen, J. J. & Skov, M. B. (2005) A Review of Research Methods in Children's Technology Design. *Proceedings of the 4th International Conference for Interaction Design and Children (IDC'04)*, 9-16, Boulder, CO).

During my work with this thesis I have encountered many skilled and accomplished people who have influenced the process and course of the Ph.D. First and foremost I owe great thanks to my colleagues in the HCI group, Jan Stage, Mikael B. Skov, Jesper Kjeldskov, Rune Thaarup Høegh, Jeni Paay and Benedikte Skibsted Als for the many

enlightening discussions, but also my other colleagues in the Information Systems Unit, the people of the Systems Development group, Ivan Aaen, Peter Dolog, Fred Durao, Lise Tordrup Hermansen, Jens Henrik Hosbond, Karsten Jahn, Annette Moss, Andreas Munk-Madsen, Peter Axel Nielsen, John Persson, Jeremy Rose, Gitte Bay Tjørnehøj and Lin Yujian as well as the secretaries Ulla Langballe and Helle Schroll have contributed with their knowledge and guidance.

This thesis would not have been possible without the funding from the USE-project, and I therefore which to extend my thanks to the participants of the project. Firstly my colleagues, Rune Thaarup Høegh, Mikael B. Skov and Jan Stage (Aalborg University) and Erik Frøkjær, Kasper Hornbæk, Mie Nørgaard and Tobias Uldall-Espersen (University of Copenhagen). Secondly, I also wish to thank my collaboration partners in the project, Lyngsoe Systems A/S, the home healthcare workers of Aars Municipality and the sixty 7th-grade students from five schools in the Aalborg area for lending their time to the project. In connection to this, I wish to thank my co-authors of the papers presented in this thesis, Benedikte Skibsted Als, Rune Thaarup Høegh and Mikael B. Skov for continuous fruitful collaboration and valuable discussions.

A special thank you goes to my supervisor Mikael B. Skov for continuous support and constructive criticism on research papers and during the writing of this thesis as well as for his encouragement to apply for a Ph.D. scholarship and pursue a research career.

My years as a Ph.D. student also contained a three months stay in wonderful Northern Italy, where I had been welcomed at University of Udine by Prof. Luca Chittaro and his colleagues. It was a great stay, both professionally and personally, and I wish to thank all at the HCI-Lab, Stefano Burigat, Fabio Buttussi, Luca Chittaro, Demis Corvaglia, Luca De Marco, Lucio Ieronutti, Daniele Nadalutti, Roberto Ranon, Augusto Senerchia and Alessandro Verona. Also I thank my colleagues in the projects Indtal, MAGNET and CNTK at the Department of Electronic Systems Ove Andersen, Tom Brøndsted, Paul Dalsgaard, Haitian Xu, Kasper Løvborg Jensen, Lars Bo Larsen, Børge Lindberg, Christian Fischer Pedersen, Jakob Schou Pedersen and Zheng-Hua Tan.

I could not have accomplished this without the continuous support of my family: My mother and father, Lis & Poul Jul Sørensen, for their lifelong support and faith in me, and my sister and brother, Rikke and Jacob Jul Sørensen, for supplying the occasionally much needed social breaks. Also a loving thought to my late mother-in-law who was always so proud of my achievements. And finally, to my dearest husband, Michael Jul Jensen, who has encouraged and supported me unconditionally

and with great devotion every step of the way through the past three years, and to my son Hjalte Jul Jensen for always brightening up my life.

Thank you.

Janne Jul Jensen
Aalborg, April 2009

1 Introduction

Context plays an important role in usability evaluations. Usability evaluations help reveal the possible future problems of use in a system, through the involvement of potential users, solving realistic tasks (Preece et al., 1994). Good usability has become a competitive factor in many products today and the process of ensuring this has become an integrated part of the development process often in the form of a usability evaluation (Rubin, 1994). A usability evaluation is an evaluation of an application, usually in an artificial recreation of the applications use context. Therefore, the important and influential aspects of the normal use context should be recreated in the artificial context of the evaluation (Bevan & Macleod, 1994). However, even when this is taken into consideration, it is still unclear if and how this artificially created context influences the use differently than the normal use context would.

To understand what aspects of context to recreate during a usability evaluation, it is necessary to examine what context is and which aspects it contains. Dey and Abowd (2000) found that while several different understandings and definitions of context exist, most agree that context includes physical location. Context is not limited to physical location though, it can also include such aspects as e.g. cultural context (Hillier, 2003; Nivala & Sarjakoski, 2003), organisational context (Maguire, 2001; Bevan & Macleod, 1994), technological context (Jones & Marsden, 2006; Maguire, 2001) or social context (Jones & Marsden, 2006; Maguire, 2001).

This thesis deals with the social context of usability evaluations. Lacking a generally accepted definition of social context, a tentative understanding is that *the social context in relations to a user in a usability evaluation comprises people surrounding the user during evaluation and their relationship with the user.* In the remainder of this chapter I will elaborate further on the concepts usability evaluations and social context and I will finish the chapter by motivating and presenting the research questions of the thesis.

1.1 Usability Evaluations

Usability evaluation is an important part of today's software development process as it can help improve the usability of systems under development. Usability evaluations can save money, time and effort if introduced into the process correctly and at the right time (Nielsen, 1993). The justifying examples are many, but as stated by Rauterberg (2003) and Bias and Mayhew (1994) the common conclusion is not *if* usability is cost-justifiable, but rather by how much.

A usability evaluation is a process that helps identify possible weaknesses with regards to a system's usability through the involvement of actual users. This is done by having them use the system to help them solve a set of tasks that represent the future use of the system. The results of a usability evaluation can be presented in different forms, such as task completion time, subjective workload measurements, error rate or usability problems. The latter is a frequently used representation of the results of a usability evaluation (Nielsen, 1993; Wixon, 2003).

Usability evaluations involve a number of activities, e.g. designing tasks that reflect the future use of the system (Heim, 2008; Rubin, 1994), deciding on a method or protocol to be used for the evaluation (Dix et al., 2004, Schneiderman, 1998), deciding what data to collect and how to collect it (Dix et al., 2004, Preece et al., 1994), the activity of recruiting participants that are representative of the end-user group (Dix et al., 2004, Preece et al., 2007) and deciding if the evaluation is best done in a usability laboratory or as a field study (Dix et al., 2004, Heim, 2008).

One of the more predominant discussions is the choice of location for usability evaluations. Typically, the choice is between evaluating in an artificial setting such as a laboratory or in a more natural setting through a field evaluation. However, each of these settings has strengths and weaknesses (Preece et al., 2007; Markopoulos et al., 2008). An artificial setting supports control but lacks realism (Dix et al., 2004; Leventhal & Barnes, 2008; Markopoulos et al., 2008; Heim, 2008; Rubin, 1994), whereas a natural setting supplies realism but makes control more difficult (Preece et al., 2007; Rubin, 1994; Markopoulos et al., 2008; Kjeldskov & Skov, 2003). For each of the settings the aim is often to benefit from the strengths while minimising the weaknesses. In an artificial setting this is done through a simulation of context which means recreating relevant aspects of the use context to the extent possible (Bevan & Macleod, 1994; Kjeldskov & Skov, 2007; Kjeldskov & Stage, 2004), whereas in a natural setting, advanced technology increasing unobtrusiveness while maintaining control is often utilised (Schneiderman, 1998; Preece et al., 2007).

This focus on location for usability evaluations indicates that this aspect of context is considered important, when choosing which aspects of the use context to recreate in usability evaluations. Many definitions on context exist but they differ on content (Dey & Abowd, 2000). Most include location and identity of people nearby (Schilit & Theimer, 1994; Ryan, 1997; Brown, 1997), but aspects as differing as time (Ryan, 1997; Brown, 1997), temperature (Brown, 1997) and emotional state (Dey, 1998) are included in some of the definitions. Some of the most widely quoted and applied definitions is the one by Dey and Abowd (2000) who state that context is the

information that characterize the situation of a person, object or location relevant to the interaction between user and application, and the one by Schilit and Theimer (1994) stating that context consists of technical environment, user environment and physical environment. Thus, even though the definitions vary, both of them include some form of physical, technological and social context as aspects of use context.

However, despite most definitions of context containing other aspects than location, many research papers recreating use context for usability evaluations still view use context almost purely as the physical and are therefore mainly concerned with recreating this aspect during usability evaluations (Kjeldskov & Stage, 2003; Kjeldskov & Stage, 2004; Nielsen et al., 2006; Po et al., 2004). Others acknowledge social context as being part of a use context but does not report any results regarding how social context influences, or what it contributes to the outcome of an evaluation compared to other aspects of context (Bevan & Macleod, 1994; Chen & Kotz, 2000; Brooke, 1996).

1.2 Social Context in Usability Evaluations

While previous research studies in usability evaluations have largely focused on the physical aspects of context in usability evaluations, Jones and Marsden (2006) state that the social context in a usability evaluation can be equally important. One of the purposes of creating a social context in usability evaluations is to facilitate effective and efficient evaluations. Usually, the primary focus is on enabling participants to successfully think-aloud as think-aloud has been found to be rather challenging and difficult (Ericsson & Simon, 1993; Nielsen, 1993). Creating a proper social context can potentially diminish some of these challenges.

Introducing a test leader in usability evaluations can be viewed as an attempt to create a social context for the participant by having the test leader sitting next to the participant during the usability evaluation. Rubin (1994) claims that this setup enables test leaders to catch more details of the participant's interaction with the system, help participants during the usability evaluation, and making participants feel less alone. Thus, under the right circumstances a good test leader can create a social context that enables effective thinking-aloud. However, a test leader also introduces a number of possible pitfalls in a usability evaluation. Nielsen (1993) claims that a test leader influences the product of the usability evaluation by impacting the number of identified usability problems. Especially the test leader's knowledge and experience with the system being evaluated potentially influences the identification of usability problems (Nielsen, 1993). Furthermore, test leaders sometimes tend to lead rather than

enable, jump to conclusions or act too knowledgeable (Rubin, 1994), and thereby influence the process as well as the product of the usability evaluation. These pitfalls can be minimized only through increased experience as a test leader (Dumas & Loring, 2008). Increased experience will also help the test leader balancing the roles of host, leader and observer throughout the evaluation (Dumas & Loring, 2008). van den Haak and de Jong (2005) found that the presence of a test leader affects the behaviour of think-aloud participants evaluating alone as they display a heightened awareness of the test leader during the process of a usability evaluation.

The imbalanced power structure between a participant and the test leader can create undesirable social contexts as illustrated above. Involving more than one participant in each session, i.e. having peer participants collaborate while interacting with the system, can address this issue. Several research studies have investigated peer-participants collaborating through the method constructive interaction, originally introduced by O'Malley, Draper & Riley (1984), in which participants evaluate in pairs instead of the classical think-aloud protocol where participants evaluate alone (Kahler, 2000; Gutwin & Greenberg, 2000; Wildman, 1995; Wilson & Blostein, 1998). The fundamental idea is that constructive interaction inherently makes participants think-aloud as they collaborate while solving tasks in the system due to the natural dialogue that arises between two participants collaborating (O'Malley, Draper & Riley, 1984; Nielsen, 1993). Kahler (2000) confirmed this as he found that constructive interaction sparked a lively, natural, and informative conversation between participants thus affecting the process of a usability evaluation. Other researchers too argue in favour of constructive interaction as a more natural way for the participants to verbalize during an evaluation (Wildman, 1995; Wilson and Blostein, 1998), but typically with no or limited empirical evidence. Nielsen (1993) claims that constructive interaction is especially suited for usability evaluations with children as it facilitates children's verbalisation better than the classical think-aloud protocol, an assumption confirmed by Hanna et al. (1997) and Markopoulos et al. (2008). Introducing more child participants in the usability evaluation have produced heightened enjoyment of the participating children (Markopoulos and Bekker, 2003; Markopoulos et al., 2008; Höysniemi et al., 2003).

The fundamental significance and contribution of involving more participants in the same session has been questioned by a number of research studies. When evaluating a system for computer supported collaborative work, Gutwin and Greenberg (2000) found that the usability problems identified based on constructive interaction originated from a poor support of the basic collaborative work rather than

the change in social context. Furthermore, Markopoulos and Bekker (2003) found that while constructive interaction influenced the participants' enjoyment and experience of the usability evaluation, it had only limited effect on the number of identified usability problems when evaluating with children. Perhaps due to the recommendations by Nielsen (1993), Hanna et al. (1997) and van Kesteren et al. (2003) many evaluation studies with children employ constructive interaction (Montemayor et al., 2002; Benford et al., 2000; Danesh et al., 2001) but typically they provide no empirical reflections on its use. Thus, while introducing more peer participants in usability evaluation sessions and thereby changing the social context for the evaluation, we still have only limited understanding of how and why participants interact and influence each other and how this impacts the process and product of the usability evaluation.

The social context of a usability evaluation consists, however, not always solely of the people actively involved in the execution of the usability evaluation, such as participants and test leader. Sometimes passive additional parties will be present as part of the attempt to simulate a realistic social context of the usability evaluation. Rubin (1994) introduces additional testing role participants needed to simulate different roles during the evaluation as part of the design of the evaluation, for instance to staff a hotline, reply to an e-mail or act as a colleague. Still, in most cases there is little empirical data clarifying how this presence of passive additional parties affects the process and product of the usability evaluation (Kjeldskov et al., 2004; Bekker et al., 2003; Bers et al., 1998).

In summary, the importance of social context has been stated by several usability handbooks (Jones & Marsden, 2006; Frohlich & Kraut, 2003; Nielsen, 1993), yet little research has attempted to provide an overall understanding of the effects of social context on the process and product of usability evaluations. Present research has examined elements of social context including the role of the test leader, the inclusion of multiple participants and the inclusion of passive additional parties in an attempt to understand the influence of social context in usability evaluations. However, there is little coherence in the understanding of social context, and there is a general disagreement on how social context impacts usability evaluations. Thus the research on social context is scattered and scarce, lacking a unifying overview. Therefore an understanding of the key characteristics of social context and how it impacts the process and product of a usability evaluation is needed.

As part of achieving such an understanding, three issues need to be further addressed: Firstly, I will explore the key characteristics of social context in usability evaluations. Secondly, I will explore how social context affects verbalisation and

collaboration in usability evaluations. Thirdly, I will examine how social context influence problem identification in usability evaluations. These three issues form the basis of the research questions of my thesis.

1.3 Research Questions

Based on my discussion of social context in usability evaluations I present the three research questions of my thesis:

1.3.1 Concepts

The exploration of social context showed that a generally accepted definition of social context does not exist. Furthermore, social context is a difficult concept to grasp (Jones & Marsden, 2006; Nivala & Sarjakoski, 2003). This leads to the first sub-question of my thesis:

---

a. Which key concepts characterise social context in usability evaluations?

This question is answered through the introduction of the theory of behaviour settings from environmental psychology which provides tools for understanding human behaviour as well as through a review of current practice with regards to social context in usability evaluations involving children.

1.3.2 Processes

Social context has been found to impact usability evaluations, but it is still unclear how this will affect the process of the evaluation with regards to the collaboration and verbalisation of the participants. This leads to the second sub-question of my thesis:

---

b. How does social context affect collaboration and verbalisation in usability evaluations?

This question is addressed through three papers: Two papers that report on experiments designed to clarify the influence of social context on the process of usability evaluations and one paper that provides an overview of literature concerned with social context in the process of usability evaluations involving children.

1.3.3 Products

The researchers reporting on findings regarding social context, rarely state explicitly how this has affected the outcome of the evaluation. Thus it is unclear if more problems, less problems or just different problems are discovered due to the simulation of social context in usability evaluations. This leads to the third sub-question of my thesis:

---

c. How can social context support problem identification in usability evaluations?

This question is addressed through three papers: Two papers that report on experiments designed to clarify the influence of social context on the product of usability evaluations and one paper that provides an overview of literature concerned with the impact of social context on the product of usability evaluations involving children.

The first question will be addressed in chapter 2 and paper contribution 5. The second and third questions are addressed through all five paper contributions summarised and presented in chapter 3.

2 Social Context in Usability Evaluations

In this chapter I will introduce a theory to understand social context in usability evaluations. A limited amount of research seems to consider social context in usability evaluations and there is no generally accepted understanding of the concept of social context in usability evaluations. I adapt the theory of behaviour settings as a way to understand and describe social context in usability evaluations. The rational behind this choice is twofold: Behaviour settings is a theory that provides a well-founded and powerful theoretical framework for understanding social context. Furthermore it has previously been adapted and used by Blanchard (2004) as a way to understand and explain virtual communities.

2.1 Understanding Social Context through Behaviour Settings

The concept of behaviour settings was introduced by Roger Garlock Barker in the late 1940s as stated by Schoggen (1989). He continuously collected empirical data from a small town of less than 2000 people from 1947 through 1972 based on which he developed the theory of behaviour settings. His reason for developing this theory was, in his own words:

> *"The physical sciences have avoided phenomena with behavior as a component, and the behavioral sciences have avoided phenomena with physical things and conditions as elements. (…) We lack science of things and occurrences that have both physical and behavioral attributes. Behavior settings are such phenomena (…)"*
>
> (Barker, 1978, p19)

*Behaviour setting*s consist of two elements, *behaviour* and *milieu* (setting). Behaviour comprises the way the people occupying the behaviour setting act towards each other. Milieu is a combination of time, place and things and the milieu of a behaviour setting also exist outside of the behaviour setting (Barker, 1978). As an example, a university lecture can be considered a behaviour setting, where the behaviour is the way the students and the lecturer is expected to act during such a lecture (the lecturer speaking, the students listening and taking notes, students sitting down facing the lecturer and being quiet, a serious and quiet mood) and milieu is the actual auditorium and the table, chairs and AV equipment in that auditorium. The students and the lecturer, as well as the auditorium and its content will exist when the university lecture is not in

progress, but only when combined will they make up the behaviour setting named university lecture.

Wicker (1992) agrees with the components of behaviour settings presented by Barker, when he states that behaviour settings are well defined small social systems consisting of *people* and *objects* and confined by *time* and *place*. The objects and the way people interact with them is what make a behaviour setting, while the time and place informs of the temporal and physical boundaries of the behaviour setting (Blanchard, 2004). The primary of the four components is people, since behaviour settings only exist when occupied by at least one person (Barker, 1978) thus, what makes the university lecture a behaviour setting is the combination of milieu with the presence of a lecturer and a number of students.

The milieu has two distinct features. It is *circumjacent* and *synomorphic* to the behaviour. Circumjacent means surrounding without a break in time or space (Barker, 1978), e.g. the university lecture begins at 8:00 and ends at 10:00 and does not leave the auditorium in that period of time. Synomorphic means similar in structure. Often the physical boundaries of a behaviour setting are similar to the boundaries of the behaviour setting (Barker, 1978), e.g. the walls of the auditorium are the physical boundaries of the milieu, but it is also the boundaries of the lecture that takes place in that auditorium. Similarly internally in the behaviour setting, the objects are structured to fit the behaviour setting, as in the chairs of the auditorium (objects) face the blackboards and teachers desk (objects), in the same way that the students (behaviour) face the lecturer (behaviour). It is usually the people of the behaviour setting that arrange the objects to fit the behaviour, and they are then also constrained by this (Wicker, 1987). E.g. the arrangement of objects in the auditorium facilitates lecturing, but makes discussions between students more difficult. Because of this, behaviour and milieu are called *synomorphs*.

Although there is a general agreement that behaviour settings do not exists without at least one person, opinions differ when it comes to the individuality of the occupants. Barker (1978) states that the occupants of a given behaviour setting can be substituted with other individuals without this substitution affecting the behaviour setting as a whole. He gives an example of a fourth grade class, where every year not only the students, but also the teacher are substituted with new students and a new teacher, but even so, the behaviour setting fourth grade exists unchanged. This has been disputed by Wicker (1987) and others who claim that this view is too strict. To exemplify this opinion, the fourth grade from before does overall stay the same compared to the year before, but on a more detailed level, the new teacher might apply a slightly different style of teaching and the children might have a differing skill set

compared to the children the year before. A similar situation exists for usability evaluations. If the exact same evaluation is conducted twice with two different groups of representative users, it is unlikely that the two evaluations will produce the exact same set of usability problems. Thus in this case, the experience supports the viewpoint of Wicker.

Barker also explored how behaviour settings affected the behaviour of its occupants. The occupants could be the same across multiple behaviour settings and would express varying behaviour, depending on the setting (1978). Both Barker (1978) and Wicker (1979) write about the setting program of a behaviour setting, which is the way the settings occupants are expected to behave in the setting. An example of this could be the behaviour patterns in a day of the life of the students from before. Here the same students occupy three different behaviour settings and display three different types of behaviour caused by the setting program of the behaviour setting:

- **University Lecture:** Organised activity, little change in position, serious mood, limited variety of behaviour, mainly sitting, reading, writing and listening.
- **Lunch break:** Partly organised activity, mostly seated position, light and cheerful mood, main activity is eating and talking, but organising, walking and other activities can take place.
- **Social activities:** Unorganised activity, varied positions, exuberant mood, wide variety of activities, with talking and laughing being predominant.

Since this thesis is specifically concerned with the social context of usability evaluations and especially the power of the people involved, I will focus on Barkers understanding of the behaviour aspect of behaviour settings, since this is behaviour settings' pendant to the social context of usability evaluations.

An important feature of behaviour settings is the power that a given occupant exercises over the behaviour setting they occupy. Barker and Schoggen (1978) focus on the roles of the inhabitants of the behaviour setting and how they exercise power over the setting or parts of it. The roles are divided into seven levels of power: *Potentials* possess the least power and is very peripheral to the setting (although this can potentially change) and *single leader* is the most powerful and central role. The roles and their attributes are listed in table 1.

Each behaviour setting has positions to be occupied by human components. These are called *habitat claims* and can be explained as the role a person plays or the job a person holds in a behaviour setting. In the university lecture example, the lecturer position is an example of a habitat claim and requires a human component that

possesses the right skills and knowledge for the claim. Barker and Schoggen (1978) divide habitat claims into two categories. *Single leader*, *multiple leaders* and *joint leaders* are named *operatives* by Barker and Schoggen (1978) and comprise the human components that have direct control of the setting. The other category has not been named in a similar fashion by Barker and Schoggen. This group consists of the human components that have indirect or no control over the setting and therefore I will name this category *non-operatives* throughout this thesis. They include *members*, *spectators*, *neutrals* and *potentials*.

Habitat Claims	Level of Power	Definition
Operatives	Single leader	Direct control of entire setting
	Multiple leaders	Direct, shared control of entire setting
	Joint leaders	Direct, shared control of part of setting
Non-operatives	Members	Indirect control of most of setting
	Spectators	Some influence on part of setting
	Neutrals	Almost no power
	Potentials	Potential inhabitants of the setting

Table 1: The different roles and levels of power that people of a behaviour setting can possess (adapted from Barker and Schoggen, 1978).

The operatives are particularly important in a behaviour setting since they operate the setting program as well as maintain the structural unit of the setting. This importance is double sided since it not only entails power over both the setting and its inhabitants but also responsibility. The setting cannot exist without them and all important and difficult actions are carried out by them. The setting is controlled by them but they are also controlled by the setting (Barker and Schoggen, 1978). The roles of the non-operatives, however, are also important. As an example, imagine a university lecture that attracts no students over a period of time.

I will illustrate habitat claims through the example of the university lecture. Traditionally the lecturer would be the single leader of the university lecture setting, but it is possible that he has a teaching assistant assigned to the course as well. This assistant would then be a joint leader. Or some courses are taught by multiple lecturers that would then be multiple leaders. The students can also be divided up into groups: The actively participating students would have the member's role, whereas more passive students would have the spectator's role. Students that by mistake have entered the wrong lecture would have neutral power and students, who are aware that the lecture takes place but have yet to attend it, have potential power that can be realised if they decide to join as either active or passive students.

I will apply the operatives and non-operatives of behaviour from behaviour setting to understand and describe social context. Applying behaviour settings in context of information technology has previously been done successfully by Blanchard (2004) who in her paper applies the theory to virtual communities in order to better understand how these function. Blanchard claims that virtual communities are becoming increasingly widespread and they come in multiple forms, attract many different types of people and are used for differing purposes, yet up until her paper there has been no actual theory developed about virtual communities. Blanchard (2004) attempts to do this by introducing the theory of behaviour settings and modifying it to account for the fundamental differences between virtual and actual communities.

2.2 Usability Evaluations

Having introduced a theory to help gain an understanding of social context in usability evaluation, I will discuss what characterise usability evaluations next. Several handbooks offer an understanding of what a usability evaluation is. Rubin (1994) states that a usability evaluation is: *A process that employs participants who are representative of the target population to evaluate the degree to which a product meets specific usability criteria.* Other handbooks may not write an actual definition, but they do agree on a list of characteristics that are typical for a usability evaluation. These include planning the evaluation process (including choice of method), involving real users, making realistic tasks, record or observe the participant throughout the evaluation and analyse the data (Dumas & Redish, 1999; Nielsen, 1993; Preece et al., 2007; Dix et al., 2004; Preece et al., 1994; Schneiderman, 1998) and concur with the definition presented by Rubin (1994). Literature largely agrees on the understanding of what a usability evaluation comprises and any disagreements are minor. Thus I choose to adopt the previously introduced definition by Rubin (1994) as my understanding of usability evaluations.

In this thesis, I will divide a usability evaluation into two parts: process and product inspired by Rubin (1994). I define the process of a usability evaluation to encompass the verbalisation and collaboration taking place between all parties participating (Nielsen, 1993; Rubin, 1994). The result of a usability evaluation can take many forms depending on the goal of the evaluation. The classical outcome of a usability evaluation is a list of usability problems (Nielsen, 1993; Wixon, 2003) and I will adopt this as my understanding of product.

2.3 Framework

The above description of social context and of usability evaluations provides the foundation for the framework below (see table 2). Each of the quadrants will be explored through my research contributions (see chapter 3). The first quadrant will

explore how the process of usability evaluations is affected by the roles played by the operatives participating. Second quadrant explores how the same process is affected by the non-operatives participating. The third quadrant will explore how the product of a usability evaluation is affected by the operatives participating and finally the fourth quadrant will explore how the product of a usability evaluation is affected by the non-operatives participating. Additionally, a fifth paper will present a review of the research done to produce an overview of current practice. All five paper contributions will be presented in the following chapter.

		Social Context	
		Operatives	**Non-operatives**
Usability Evaluation	**Process**	1	2
	Product	3	4

Table 2: The framework for the research of my thesis.

3 Research Contributions

In this chapter I present a summary for each of the five published paper contributions of this thesis. Each of the published papers has been placed in the framework according to their primary area of contribution, although they may also contribute to other areas of the framework. My fifth paper contribution is placed in the middle of the matrix as it contributes to all four quadrants (see table 3).

		Social context	
		Operative participants	**Non-operative participants**
Usability Evaluation	**Process**	1. Exploring Verbalization and Collaboration of Constructive Interaction with Children	2. A Case Study of Three Software Projects: Can Software Developers Anticipate the Usability Problems in their Software?
		5. A Classification of Research Methods and Purposes in Child-Computer Interaction	
	Product	3. Composing Children Dyads in Constructive Interaction: A Comparison of Usability Testing Methods for Problem Identification.	4. Evaluating in a Healthcare Setting: A Comparison between Concurrent and Retrospective Verbalisation

Table 3: The five published papers spanning the framework of my research.

3.1 Contribution 1

Exploring Verbalization and Collaboration of Constructive Interaction with Children

Als, B. S., Jensen, J. J. & Skov, M. B. (2005) Exploring Verbalization and Collaboration of Constructive Interaction with Children. *Proceedings of the 10th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'05)*, 443-456, Berlin: Springer-Verlag.

This paper reports on an experiment exploring how the process of usability evaluations is affected by changes in the social context, in this case the composition of operatives

included in the usability evaluation. Previous research has claimed that children tend to find it difficult to verbalise applying the classical think-aloud protocol, due to the unnaturalness of the situation (Hanna et al., 1997; Markopoulos et al., 2008). To address this problem constructive interaction has been suggested as supporting verbalisation in a more natural manner during usability evaluations with children (Markopoulos et al., 2008; Nielsen, 1993). Our experiment explored think aloud with one participant (joint leader) and constructive interaction with two participants (multiple leaders), and as an extra dimension during constructive interaction, we varied the acquaintance of the multiple leaders. The following setup was applied: 60 children evaluated a mobile system through a set of tasks in three different laboratory setups. 12 children evaluated the system using a standard think aloud protocol, 24 children evaluated the system in acquainted pairs applying constructive interaction and the final 24 children evaluated the system in unacquainted pairs also applying constructive interaction. Each of the setups included an equal number of boys/pairs of boys and girls/pairs of girls to reduce gender bias. After each evaluation session the participants were subjected to a NASA TLX test to measure their mental workload during the process. All sessions were videotaped and analysed afterwards focusing on the process of the usability evaluation. The process of the usability evaluation included the test leader's influence and interaction with the participants as well as the children's ability to collaborate on the tasks, their ability to verbalise and how the different social contexts affected their performance, experience of the evaluation and workload.

The results of our experiment indicate that the process of a usability evaluation is affected by the acquaintance of the multiple leaders involved. Although the level of verbalisation was higher using constructive interaction, the process did not seem to benefit from applying constructive interaction as has been suggested by literature (Nielsen, 1993; Markopoulos et al., 2008), since often the operatives would talk aloud instead of think aloud, hence they would verbalise their actions rather than the thoughts behind those actions. The collaboration during the evaluation is also affected by the configuration of the operatives since acquainted multiple leaders tend to find the evaluation less demanding and exhibit a greater satisfaction with their own work than the non-acquainted multiple leaders. However, the effect of the configuration of multiple leaders on the evaluation differed for acquainted boys and acquainted girls, since contrary to their own perception of the process, the acquainted girls revealed a rather poor level of collaboration, whereas the acquainted boys worked rather well together.

In conclusion, contrary to claims in literature, the social context of paired children in the process of a usability evaluation does not necessarily heighten the quality of

verbalisation. Furthermore, the social context of paired boys improved collaboration whereas the social context of paired girls diminished collaboration, and finally all dyads found the evaluation more satisfying and less demanding.

3.2 Contribution 2

A Case Study of Three Software Projects: Can Software Developers Anticipate the Usability Problems in their Software?

Høegh, R. Th. & Jensen, J. J. (2008) A Case Study of Three Software Projects: Can Software Developers Anticipate the Usability Problems in their Software? *Behaviour & Information Technology (BIT)*, 27(4) 307-312, Taylor & Francis Group.

This paper reports on how the presence of non-operatives during a usability evaluation affects the process of the usability evaluation. This was examined through a usability evaluation setup in which the developers of three software applications acted as non-operatives. They were asked to individually describe the usability problems of the application that they had developed with no interaction between participants, yet all sitting in the same room in the presence of the other participants, who would then act as non-operatives to each other. After completing the individual tasks, all participants engaged in a group discussion regarding the findings and their validity, causing all the participants to become operatives. This was then repeated with the developers being asked to rate the problems according to severity, which was then discussed in plenum too. The problems and their rating were compared to the results of a regular usability evaluation which was used for comparison, thus making this a rather unusual usability evaluation, given that the participants were only present on video clips and through a problem list.

The individual tasks of the experiment showed that influence of non-operatives in a usability evaluation was fairly vague and difficult to observe. It was noticeable that some participants seemed to be aware of the non-operatives around them and they might also have been influenced by their presence, but it is unclear how. Some appeared to become more active when registering activity in others. Thus, when another participant started to write intensely it would sometimes cause the participants sitting nearby to also increase their level of activity. One of the participants was the daily leader of the department, and it seemed like the people sitting near him during the individual tasks as well as the group discussion were slightly more engaged in the process. However, these results regarding the influence of non-operatives during individual tasks are subject to some uncertainty.

During the group discussion part of the workshop, the roles of the participants changed to operatives. The influence of the social context in this scenario was much clearer, as the discussion was very lively and the participants were very engaged in discussing the origin and validity of problems. It also seemed that some of the participants who were less active during the individual tasks engaged more in the group discussion part. The test leaders role was more subtle during this part, since the participants had no problems keeping the discussion going. This is in agreement with the findings of van den Haak and de Jong (2005) who find that the participants are much more aware of the test leader when evaluating alone than when evaluating in pairs.

In conclusion, the social context of a usability evaluation is affected by the presence of non-operatives during individual tasks although the effect of non-operatives is somewhat unclear and difficult to register. The social context is also affected by involving multiple participants in the usability evaluation, which produces a livelier and more engaging social context and makes the presence of a test leader less important.

3.3 Contribution 3

Composing Children Dyads in Constructive Interaction: A Comparison of Usability Testing Methods for Problem Identification.

Als, B. S., Jensen, J. J. & Skov, M. B. (2009) Composing Children Dyads in Constructive Interaction: A Comparison of Usability Testing Methods for Problem Identification. (Extended version of Als, B. S., Jensen, J. J. & Skov, M. B. (2005) Comparison of Think-Aloud and Constructive Interaction in Usability Testing with Children. *Proceedings of the 4th International Conference for Interaction Design and Children (IDC'05)*, 80-87, New York: ACM Press.)

This paper reports on how the product of usability evaluations is influenced by the social context of the evaluation, exemplified by the configuration of operatives involved in the usability evaluation. Previous research has claimed that children tend to find it difficult to verbalise applying the classical think-aloud protocol, due to the unnaturalness of the situation, thus making the detection of usability problems difficult. To overcome this problem constructive interaction has been suggested as better supporting verbalisation of usability problems during usability evaluations with children (Nielsen, 1993; Hanna et al., 1997). Our experiment explored think aloud with a joint leader and constructive interaction with two multiple leaders, and as an extra dimension during constructive interaction, we varied the acquaintance of the multiple

leaders. The following setup was applied: 60 children evaluating a mobile system through a set of tasks in three different laboratory setups. 12 children evaluated the system using a standard think aloud protocol, 24 children evaluated the system in 12 acquainted pairs using constructive interaction and the final 24 evaluated the system in unacquainted pairs also using constructive interaction. Each of the setups included an equal number of boys/pairs of boys and girls/pairs of girls to reduce gender bias. All sessions were videotaped and analysed afterwards. This experiment is the same as in contribution 1 but this contribution reports on a different aspect of the experiment, namely how the social context influences the product of the usability evaluation. Therefore the focus of the analysis was the product which involved determining usability problems experienced. A problem was identified as a delay of the user, an irritation to the user or a, to the user, surprising behaviour by the system. For each problem the severity was determined, depending on the length of the delay, the level of irritation and level of surprise cause by system behaviour. Also task completion time, task completion and error rate were identified.

Our findings showed that social context of a usability evaluation influences the product through the composition of multiple leaders during the usability evaluation. There were few significant differences between the product of a usability evaluation involving a joint leader and the product of a usability evaluation involving two multiple leaders in terms of number of problems experienced. However, the social context influenced the diversity of problems experienced as acquainted multiple leaders experienced more different problems of all severities, especially critical ones, than the other setups. Thus, the social relation of multiple leaders influence the diversity of problems discovered. Furthermore, the non-acquainted multiple leaders found only three cosmetic problems that were not found by either of the other two setups (acquainted multiple leaders and joint leader). Thus, non-acquainted multiple leaders present little added value to the other two setups with regards to the product of usability evaluations.

In conclusion, the product of usability evaluations is affected by the social context of the operatives. Acquainted multiple leaders find a higher diversity of problems than non-acquainted multiple leaders and a joint leaders and non-acquainted multiple leaders find almost no problems not detected by the other compositions of operatives.

3.4 Contribution 4

Evaluating in a Healthcare Setting: Comparison between Concurrent and Retrospective Verbalisation

Jensen, J. J. (2007) Evaluating in a Healthcare Setting: A Comparison between Concurrent and Retrospective Verbalisation. *Proceedings of the 12th International Conference on Human-Computer Interaction - Interaction Design and Usability*, 508-516, Berlin: Springer-Verlag.

This paper reports on a study examining how the presence of non-operatives affects the product of a usability evaluation. The experiment is a comparison study in which the effect of the presence or absence of non-operatives is compared through the use of a PDA as a supporting tool in home healthcare. A field trial was set up in which 15 home healthcare workers were asked to solve a number of tasks using an application running on a PDA. The physical context of the field trial was the home of an elderly citizen that was also cared for normally by the participating home healthcare workers. The elderly citizen was present as a non-operative throughout the usability evaluation. Half of the home healthcare workers were asked to verbalise during their task solving, and the sessions were recorded on video, which was analysed later. The other half were merely observed during their task solving and their sessions were also recorded on video and this video was then played back to them afterwards. They were then asked to verbalise while watching the video, and the verbalisation was recorded using a video camera. Thus, half the sessions involved the presence of a non-operative during verbalisation, whereas no non-operatives were present during the verbalisation of the other half of the sessions. The video of all 15 sessions was then analysed and problems were identified and categorised.

The findings showed a heightened cognitive burden with the participants having a non-operative present during the usability evaluation, resulting in the participant either focusing on task solving and forgetting to verbalise, thus experiencing few problems, or focusing on verbalising and having trouble concentrating on the task solving, thus experiencing many problems. While it is possible that this is caused by the duality of having to verbalise and task solve simultaneously, it cannot be ruled out that the presence of a non-operative may have added to this cognitive burden and therefore have affected the product of the usability evaluation.

The other half of the sessions, where no non-operatives were present during verbalisation, on average experienced noticeably fewer problems than the session with non-operatives present. This may have several origins. It may be due to not being

affected by having a non-operative present during the verbalisation, but it may also be caused by not having to verbalise and task solve simultaneously and finally the small amount of time that passes from the actual task solving to reviewing the video while verbalising may cause the participants memory to fade and problems to be forgotten.

In conclusion, the experiment had multiple research purposes, one being to examine how the product of a usability evaluation is affected by the presence or absence of non-operatives during the usability evaluation. It was unclear if the participants verbalising with a non-operative present were influenced by the awkwardness or private nature of the information they were verbalising about, but some influence could not be ruled out. Similarly, the lower amount of problems experienced in sessions with no non-operatives present could not be unequivocally attributed to the absence of non-operatives, but could not be ruled out either.

3.5 Contribution 5

A Classification of Research Methods and Purposes in Child-Computer Interaction

Jensen, J. J. & Skov, M. B. (2009) A Classification of Research Methods and Purposes in Child-Computer Interaction. (Extended version of Jensen, J. J. & Skov, M. B. (2005) A Review of Research Methods in Children's Technology Design. *Proceedings of the 4th International Conference for Interaction Design and Children (IDC'04)*, 9-16, Boulder, CO).

This paper reports on a literature review of publications within child-computer interaction. 3295 papers published in ten of the most prominent outlets on HCI and child-computer interaction in the period 1996-2005 had at least the abstract (and if necessary, introduction and more) read in order to filter out only the papers concerned with child-computer interaction. The result was 132 papers and each of these were read fully and classified by each author individually after which the final classification was negotiated collaboratively. The research papers were categorised in a two-dimensional framework originally published by Wynekoop and Conger (1990). The two dimensions are *research method* which contains eight different categories and *purpose*, which contains five different categories. The eight different categories of research method are *case studies, field studies, action research, lab experiments, survey research, applied research, basic research* and *normative writings*. The first three are methods conducted in natural settings, the fourth method is applied in an artificial setting and finally the last four are environmentally independent methods. The five categories of purpose are *understand, engineer*, *re-engineer*, *evaluation* and *description*.

The conclusions of the paper are that there is a strong emphasis on doing research in natural settings within child-computer interaction and a rather weak focus on reporting issues of understanding, but rather on engineering or evaluation. Both gender and age is reported as being important factors in child-computer interaction, yet only gender is reported as having been actively investigated in the papers.

As this thesis addresses social context in usability evaluations, I have chosen to re-categorise the papers from this review that do usability evaluations. I will categorise them according to the framework of my thesis (the distribution of the re-categorised papers in the framework can be seen in Appendix B). 87 papers from my original review had been categorised under the purpose of evaluation. Out of these, 23 papers did not report on any type of social context. These were typically papers reporting on single testers, with no test leader present or no empirical findings concerning the test leader. Regarding social context, 61 papers included operatives in some way (typically as children evaluating in pairs or groups), whereas only 10 papers included non-operatives, which supports earlier claims that non-operatives is a subject dealt with by very few researchers. Furthermore, all of these 10 papers include non-operatives solely as a remark and not as a focus of their research. The division between product and process is slightly more evened out as 34 papers report on process while 58 papers report on product. The papers in the process category typically report on findings concerning a method employed or regarding verbalisation or collaboration between participants and/or test leader, while the papers in the product category report on various findings regarding a specific application or system being evaluated. In many cases a paper would fall in more than one category.

In conclusion, only about 74% (64/87) of usability evaluation papers are also concerned with social context in some form. Out of these nearly all (95%, 61/64) involve operatives in some form, while a mere 16% (10/64) included non-operatives, and often only as a remark. The main focus in usability evaluations involving social context is product, as 91% (58/64) report on this, while 53% (34/64) report on process in their research.

4 Research Methodology and Purpose

This chapter elaborates on the research methods applied throughout my research for this thesis. To support this elaboration, I utilise the framework by Wynekoop & Conger (1990) which considers research purposes and research methods. This framework has previously been applied within HCI by Kjeldskov and Graham (2003) categorising mobile HCI as well as in my fifth paper contribution categorising child-computer interaction.

The framework by Wynekoop and Conger (1990) introduces a method for categorising research according to the purpose of the research and the method applied. They define five different research purposes (understanding, engineering, re-engineering, evaluation and description) inspired by research purpose as introduced by Basili et al. (1986). However, Wynekoop and Conger (1990) apply the framework in the field of computer aided software engineering tools in which the focus is on a specific product or application. Thus their definition of the categories in research purpose is aimed at tools. When exploring the research purpose of my research within HCI, the focus is a bit different though, as much of my research focuses on the method applied, rather than the product. However, I will keep the five research purposes as defined by Wynekoop and Conger (1990), while expanding them to include methods as the object of interest too.

In the second dimension, Wynekoop and Conger (1990) define eight research methods (case studies, field studies, action research, laboratory experiments, survey research, applied research, basic research, and normative writings) which are inspired by the method categories introduced by Scott Morton (1985). The first three research methods are characterised by taking place in a natural setting, the fourth takes place in an artificial setting while the last four research methods are environment independent settings. A more detailed description of each of the research purposes and research methods can be found in Wynekoop and Conger (1990).

Thus, inspired by the framework of Wynekoop & Conger (1990) I will structure the remainder of this chapter according to the research methods applied. For each of the applied methods I will discuss the research purpose chosen. I will then continue to discuss the actual activities including the influence of the research setting of the activities and finally reflect upon strengths and weaknesses of the choices made. This structure is presented in table 4.

Research Method	Research Purpose	Research Setting	Research Paper
Laboratory Experiment	Evaluation Understanding	Artificial	Contribution 1, 2 & 3
Field Study	Evaluation Understanding	Natural	Contribution 4
Survey Research	Understanding Description	Environment Independent	Contribution 5

Table 4: The research method applied throughout the research of this thesis and the research purpose, research setting and research paper associated with each of the methods.

4.1 Laboratory Experiments

Laboratory experiments take place in an artificial setting created by researchers with the purpose of controlling the manipulation of variables and avoiding unwanted disturbances, according to Wynekoop and Conger (1990). This setting provides a high level of control but at the expense of the realism of the setting. This is in agreement with Galliers (1992) who describe a similar approach for laboratory experiments including the same strengths and weaknesses. For the research of this thesis four laboratory experiments were conducted and documented in three of my paper contributions (Høegh & Jensen, 2008; Als et al., 2009; Als et al., 2005).

Evaluations are conducted to compare systems, to assess certain properties or to verify functionality and are in the form of a structured study (Wynekoop & Conger, 1990), while the purpose of understanding is an attempt to grasp the meaning of the object being studied. Both these purposes form the basis of the research conducted in paper contribution 1 and 3 (Als et al., 2005; Als et al., 2009). Primary purpose was evaluation of the verbalisation protocol and its effectiveness under varying compositions of participants. Secondary purpose was understanding, since the objective of the evaluation was to gain an understanding of how the composition of the dyads participating would affect the process as well as the product of usability evaluations.

Conducting a laboratory experiment to evaluate is a classical approach within HCI. The objective is typically to compare or evaluate products and methods (Schneiderman, 1998; Preece et al., 2007; Nielsen, 1993; Rubin, 1994). Laboratory evaluations contain a range of variables that can be manipulated. These include e.g. the number and type of participants involved, the verbalisation protocol utilised or the role of the test leader (Rubin, 1994; Nielsen, 1993; Preece et al., 2007). Furthermore it is characterised by the artificial setting in which the usability evaluation is taking place. This artificial setting offers a great deal of control over the setting because outside

disturbances are eliminated (Preece et al., 2007, Rubin, 1994; Markopoulos et al., 2008) and it facilitates data collection in an unobtrusive manner (Dumas & Redish, 1999; Nielsen, 1993). However, such a laboratory is rarely similar to the use context of the application being evaluated and thus the impact of the artificial context may differ from the impact of the use context (Preece et al., 2007). This may be minimised by analysing the influential parameters in the use context and then simulate these in the laboratory to the extent possible, but this is not a trivial task (Bevan & Macleod, 1994; Kjeldskov & Skov, 2007).

The experiment conducted in paper contribution 1 and 3 (Als et al., 2005; Als et al., 2009) was a laboratory evaluation in the form of a classical usability evaluation involving participants, verbalising while solving tasks and being observed by a test leader. The participants involved were children applying either think aloud or constructive interaction and video data was recorded and analysed. For this experiment it was important to gain a high level of control in order to measure only changes caused by the variables deliberately being manipulated (the composition of the participants) while keeping other variable changes and disturbances to a minimum and based on this a laboratory evaluation was chosen.

In conclusion, a classical laboratory-based evaluation offers control, but it also has weaknesses as its artificial setting offers little to no realism (Dix et al., 2004; Leventhal & Barnes, 2008; Heim, 2008; Rubin, 1994) and results gained here may therefore be difficult to generalise to a real world setting (Wynekoop & Conger, 1990). However, since the main objective of this experiment was to study aspects of usability evaluations, the weaknesses of a laboratory experiment in an artificial setting become less dominant, simply because the artificial setting of a usability evaluation can be considered the natural setting.

As established earlier, the purpose of understanding is an attempt to grasp the meaning of the object being studied. The purpose of the laboratory experiments conducted in paper contribution 2 (Høegh & Jensen, 2008) was understanding, as the objective was to explore to what extent developers are able to predict usability problems in their own software.

The laboratory experiments of paper contribution 2 (Høegh & Jensen, 2008) took place in a laboratory-like setting on site of the organisation and involved software developers. To uncover to what extent developers were aware of potential usability problems in their software, individual questionnaires were utilised, followed by collective analysis of video of actual users using the system. Three experiments were conducted in which the test leader was more active in facilitating discussions between

participants and the focus was not strictly on problem identification but also on what constitutes a problem and if/how they should be solved.

In conclusion, the weaknesses of this approach are that the limited number of participants restricts the generalisability of the results and the unique character of the experiment limits its use in other situations (Galliers, 1992).

4.2 Field Studies

Field studies take place in a natural use context of an application. According to Wynekoop and Conger (1990), a field study takes the form of either a study or an experiment. A study is characterised by being non-experimental with no manipulation of variables, recall-based and based on self reporting by the participants. Field studies thus are unobtrusive and high on realism (Wynekoop & Conger, 1990). A field experiment, on the other hand has a higher degree of control and manipulation. This has the effect of minimizing unwanted disturbances, while also lowering the realism. Galliers's (1992) description of field experiments is similar but unlike Wynekoop and Conger (1990) he does not describe field studies. Field studies are according to Wynekoop and Conger characterised by not manipulating variables, but simply observe what is. One field experiment was conducted for the research of this thesis.

Evaluations are conducted to compare systems, to assess certain properties or to verify functionality and are in the form of a structured study (Wynekoop & Conger, 1990), while the purpose of understanding is an attempt to grasp the meaning of the object being studied. My fourth paper contribution (Jensen, 2008) was based on both these purposes. The primary purpose was to compare the effectiveness of concurrent and retrospective verbalisation for verbalising during an evaluation in the field. The secondary purpose was to understand how verbalisation is affected by the presence of passive additional parties. The assumption was that the presence of passive additional parties while evaluating in the natural use context could influence the participants' inclination to verbalise.

When field studies are conducted with the purpose of evaluating, it is typically a method applied late in the development process to evaluate a product close to release (Rubin, 1994). Usually the focus of a field evaluation (evaluating through a field study) is more on the interaction with the context during use (Preece et al., 2007). Especially social context, which is more transient, can be difficult to simulate in an artificial setting and thus may be more easily obtainable through a field study (Jones & Marsden, 2006). However, a field evaluation offers less control and makes data collection difficult (Preece et al., 2007; Rubin, 1994; Kjeldskov & Skov, 2003).

For the research in paper contribution 4 (Jensen, 2007), a field evaluation was conducted in which half the participants solved tasks using an application in a natural

use context and with passive additional parties present, while verbalising concurrently. The other half of the participants also used an application in its natural use context, but did not verbalise in the presence of passive additional parties while solving the tasks. Instead the evaluation was recorded on video. This video recording was then played back to them afterwards and they would verbalise while watching themselves solve the tasks.

In conclusion, a field study supplies the realistic context needed to explore if such a context influences the participants ability to verbalise compared to verbalising in an artificial setting. However, the realism of the context was influenced by the presence of a test leader carrying equipment and the tasks not being actual work tasks, but simulated tasks produced for the evaluation. Similarly, lack of control and difficult data collection were issues that also arose in this experiment. Finally, the fact that the data used in the application during the usability evaluation was, due to security and privacy reasons, not directly linked to the passive additional parties involved in the evaluation seemed to cancel out any possible hesitance towards verbalising about otherwise personal and private data.

4.3 Survey Research

Survey research is characterised by being environment independent and by the possibility of having large sample sizes without a high resource cost (Schneiderman, 1998, Wynekoop & Conger, 1990). Due to the large sample size it lends itself well to quantitative analysis (Schneiderman, 1998). The large sample size, if properly chosen, can also reduce bias and allow for easier generalisation of results (Wynekoop & Conger, 1990). Wynekoop and Conger specifically classify literature reviews as survey research. However, this is opposed by Galliers (1992) who classifies literature reviews as descriptive or interpretive research rather than survey research. This is based on the influence that the researchers presuppositions has on the interpretation of the body of work. Despite this difference in classification, however, they largely agree on the activities as well as the strengths and weaknesses of the approach. Survey research was applied once in this thesis in the form of a literature review.

Understanding, is an attempt to grasp the meaning of the object being studied and description usually defines or describes features of ideal instances of the object studied, according to Wynekoop and Conger (1990) The primary purpose of my fifth paper contribution was understanding (Wynekoop & Conger, 1990), as the objective was to obtain an overview of the work that had been done in the field of child-computer interaction. To achieve this objective, a review survey was conducted. Secondary purpose was description (Wynekoop & Conger, 1990), as another objective

was to make recommendations for future areas of interests based on the findings of the review.

The field of child-computer interaction was chosen based on the assumption that social context has proved especially important when evaluating with children (Hanna et al., 1997; Nielsen, 1993; Markopoulos et al., 2008). For the review, the framework of Wynekoop and Conger (1990) was chosen. Ten top-level peer-reviewed journals and conferences within HCI and child-computer interaction were selected and the papers published from 1996 to 2005 in these outlets were examined for topics concerning child-computer interaction. From these outlets, 132 papers were identified as relevant to the topic. These were then read and categorised independently by two researchers.

The suitability of the framework chosen is debatable as a framework may constrain the conclusions drawn and may not offer the most appropriate categories for the literature surveyed. In the case of child-computer interaction, several papers would fall into multiple categories due to the application of adapted methods and in some cases method and purpose were not described directly and had to be interpreted, which would on occasion also prove difficult (Jensen & Skov, 2009). In other cases, the method applied was hard to identify unambiguously (Jensen & Skov, 2009). However, these difficulties are not unique to the field of child-computer interaction as they were also recognised by Kjeldskov and Graham (2003) in their application of the framework within mobile HCI. Similarly, the choice of outlets and years might not be representative and may offer a skewed snapshot of the field of research (Kjeldskov & Graham, 2003). Finally the large sample sizes often means that the richness of the data is lost, and only a few fixed aspects of each paper are being reported. On the other hand, such a classification offers an overview that would otherwise be hard to obtain and it does so at a low cost and in an environment independent setting (Wynekoop & Conger, 1990).

5 Conclusion

This thesis has addressed the influence of social context in usability evaluations. The perspective on social context has been the division into operatives and non-operatives based on the theory of behaviour settings by Roger G. Barker (1978) and usability evaluations have been characterised by process and product. This viewpoint has produced a framework to which the research conducted makes a contribution in each quadrant. In this conclusion I present the results attained throughout this thesis and the five published paper contributions. The conclusion will be structured according to the three research questions presented in chapter 1.

5.1 Research Question #1: Revisited

The first research question is concerned with the key concepts characterising social context and reads: *Which key concepts characterise social context in usability evaluations?* The findings for this research question are as follows:

1. The key concept characterising social context in usability evaluations is people. People include among others the test leader that facilitates or manages the evaluation, but can also include peer-participants (e.g. in constructive interaction). Furthermore, people can include passive additional parties who are often included to increase realism of the usability evaluation, e.g. acting as a patient in a healthcare evaluation or staffing a hotline to answer a call during the evaluation. In a literature review on child-computer interaction research, I found that 64 out of 87 papers report on or reflect upon interaction between people involved in a usability evaluation. Thus, aspects of social context seem important in evaluation studies with children. In summary, social context comprises the people surrounding the user during evaluation and their relationship with the user. Inspired by the theory of behaviour settings from environmental psychology (Barker, 1978), I divide these people into two groups according to their level of power in the usability evaluation, namely operatives and non-operatives.

2. Operatives characterise people that hold direct power over all or parts of the usability evaluation. Operatives consist of the roles of single leader, multiple leader and joint leader. Operatives usually interact actively with other operatives in the usability evaluation. For instance, a test leader can act as single leader during a usability evaluation as he holds direct power over the usability evaluation (e.g. power to stop or change the course of the evaluation). Similarly, a participant can act as joint leader due to his shared power over part of the evaluation (e.g. he can stop the

evaluation if he wishes to). Thus, the test leader (single leader) still holds more power over the evaluation than the participant (joint leader). This, however, differs in the case of constructive interaction, in which the two participants act as multiple leaders. This is based on their shared power in the evaluation. Furthermore, having two participants together with just one test leader often seems to level the playing field more. The role of the test leader is reduced to a multiple leader (equal shared power with the participants) too or even to joint leader (less active, more observing). In my literature review on child-computer interaction, my findings show that 61 out of 64 papers report involving operatives in their evaluation, usually through the application of constructive interaction or through interaction between test leader and participant(s).

3. Non-operatives hold indirect power or very limited power in a usability evaluation. Non-operatives comprise members, spectators, neutrals and potentials. Non-operatives take on a more indirect and peripheral role than operatives. For instance, a patient in the aforementioned healthcare evaluation acts as member thus holding indirect power over most of the setting. This indirect power could be utilised if for instance the patient, during the participant's interaction with him/her chose to exhibit behaviour unfitting for a patient and disruptive for the usability evaluation. If a patient is not directly in contact with the participant the patient would only act as spectator, as s/he would not have actual power but only an influence on the evaluation through e.g. disturbing behaviour. The other two non-operative roles are mostly seen in usability evaluations in the field. The social context of a field evaluation may contain people that are not directly linked to the usability evaluation. These are categorised as neutrals or potentials. Neutrals are often in the form of onlookers or bystanders, who merely happen to be present during the evaluation. Potentials play a slightly different role, as they are not yet a part of the usability evaluation, but may choose to become part of it. For instance, a nurse, who knows the evaluation is taking place, may potentially choose to engage the participant during the evaluation, thus switching role to either member or spectator. A mere 10 papers out of 64 state the presence of non-operatives in my review of child-computer interaction. Furthermore, all of these papers only mention the presence of non-operatives as a remark and do not report any results related to their presence.

5.2 Research Question #2: Revisited

The second research question addresses how social context affects verbalisation and collaboration in usability evaluations and states: *How does social context affect collaboration and verbalisation in usability evaluations?* The results for research question 2 are:

1. Pairing peer-participants successfully as multiple leaders can increase verbalisation. My findings show that the verbalisation of a usability evaluation is increased by involving children as peer-participants. Pairing children from the same class in school to collaborate as acquainted multiple leaders during a usability evaluation significantly increases verbalisation, compared to pairing children from different schools (non-acquainted multiple leaders) or involving single testers (joint leaders). Being multiple leaders presuppose that they possess equal shared power and this aspect of the role may be more readily fulfilled by peer-participants that are acquainted, thus causing this significant difference in verbalisation.

2. A lack of acquaintance can decrease collaboration between peer-participants acting as multiple leaders. If a social context is chosen carefully, it can improve collaboration between peer-participants acting as multiple leaders. However, my findings show collaboration break downs caused by a less ideal social context: Non-acquainted multiple leaders might in some cases switch to turn-taking in the task solving. This may appear due to an unwillingness (typically due to shyness) to engage in cooperation with a stranger. The turn-taking can also be a result of politeness and the fact that many children have been taught that it is polite to share. This results in a situation in effect resembling a sequential single tester session, thus not benefitting from the claimed advantages offered by constructive interaction. Some studies in my review report having involved children from multiple schools, thus presumably unacquainted, but none of them report any findings regarding their collaboration.

3. Too close acquaintance can decrease collaboration between peer participants in the roles of multiple leaders. Multiple leaders who are not only acquainted, but rather best friends display a different kind of collaboration break down. In this case, politeness is absent and they tease each other, grab the application from each others hands and obstruct the other persons work by pressing buttons in the middle of their task solving effort. This typically appears as a manifestation of their disagreement on the course of the task solving or as a demonstration by one participant experiencing the other participant as monopolising the application. Thus to avoid these types of break downs in collaboration between multiple leaders, it is beneficial to pair participants that are not close friends. The findings of my review on this matter are unclear, as some stress the importance of the peer-participants being friends, while others do not consider the level of acquaintance.

4. Due to lack of verbalisation, a test leader may dynamically have to switch role during usability evaluations, thus changing his level of power. In constructive interaction the test leader usually plays the role of joint leader (passively observing making occasional prompts), while the peer-participants act as multiple leaders (equal). However, in some cases the test leader has to actively switch role to single leader (by starting to ask questions) during the evaluation, thus making the participants joint leaders (reducing their power), simply because the verbalisation i.e. the communication between the participants does not flow as naturally as expected. It is possible that the usability evaluation may not progress properly in these situations, if the test leader does not actively switch to a more powerful role than originally anticipated. Thus, in such a case the ability to verbalise affects the process through a dynamic shift in the roles occupied by the involved parties. Therefore, to facilitate an effective usability evaluation session, a test leader should always be prepared to switch role if the situation requires it.

5. The influence of non-operatives during a usability evaluation is vague and difficult to observe. Participants working alone in the presence of other participants (thus acting as each others non-operatives) performing similar tasks display an awareness of each other and each others activities. However, it is unclear if this awareness causes changes in their behaviour. One observation shows, though, that intense writing from one non-operative can draw the attention of others nearby and sometimes seems to cause them to increase their activity too. Similarly, the presence of a senior employee seems to cause people nearby to engage slightly more in the process, suggesting that non-operatives with seniority increase the level of power in relation to people nearby. Thus, the level of power of non-operatives in a usability evaluation in relation to the participant may be affected by the level of power they hold in daily life compared to the participant. However, whether these changes in behaviour are in fact linked to the presence of non-operatives is unsure.

5.3 Research Question #3: Revisited

The third research question is concerned with how social context supports problem identification and reads: *How can social context support problem identification in usability evaluations?* The findings of my last research question are:

1. Social context can support the identification of a higher number of usability problems as well as a higher number of unique usability problems. Compared to non-acquainted multiple leaders and single testers (joint leaders), our results show that acquainted children being peer-participants (multiple leaders) in a usability evaluation

identify the highest number of usability problems of the three setups. Furthermore, they identify more unique usability problems than non-acquainted multiple leaders or single testers. The additional problems identified seem to be primarily critical or serious problems. Similarly, half of the unique problems identified by acquainted multiple leaders are also categorised as critical or serious. Thus, to identify the largest number of severe usability problems, the results indicate that acquainted multiple leaders should be chosen for participation.

2. Acquainted multiple leaders and single testers in conjunction identify the widest range of usability problems. Compared to the list of usability problems identified in total by the three setups only a few cosmetic problems are not revealed by either acquainted multiple leaders or single testers. Furthermore, all critical or serious problems are found by either acquainted multiple leaders or single testers. Thus, the range of problems can be covered by only involving single testers and acquainted dyads, and not many new problems are identified from also involving non-acquainted dyads. This may be because the coupling of acquainted multiple leaders and single testers represent the most variety in social context, and thereby covers the largest impact on problem identification.

3. Problem identification is distributed differently between different user groups verbalising. Professional adults who are asked to verbalise while task solving seems to concentrate their efforts on either verbalising or on task solving. Those participants focusing on task solving and forgetting to verbalise, experience few problems (6-11), while those focusing on verbalisation have trouble task solving, thus experiencing many problems (21-36). A similar pattern for children was not found, which indicates that this is not caused by the mental burden of verbalising concurrently. It is possible that the difference originates from variations in the distribution of power during the usability evaluation. An adult test leader with a child participant inherently, through the adult-child relationship, holds a higher level of power, which reinforces the roles of single leader and joint leader. In the case of a professional adult participant, this is not as distinct, as the participant often has an area of expertise (e.g. their area of work) that is less familiar to the test leader. The participant can therefore claim equal level of power, as both test leader and participant possess an area of expertise (usability evaluations vs. area of professional occupation) and based on this more equal distribution of power, they tend to take the roles of multiple leaders, rather than single leader and joint leader. Thus, the influence of the test leader differs, due to the different roles of the test leader.

4. Concurrent verbalisation identifies more usability problems than retrospective verbalisation. Verbalisation can be done in concurrence with task solving in the actual usability evaluation, or the usability evaluation can be silent, followed by retrospectively verbalising, while watching a playback of the usability evaluation. This changes the social context of verbalising as concurrent verbalisation takes place in the actual context of the usability evaluation, while retrospective verbalisation typically takes place afterwards in an artificial context like an office or meeting room. Verbalising concurrently in the presence of non-operatives identifies more usability problems than verbalising retrospectively without non-operatives present. The presence or absence of a non-operative represents a change in the social context and it is unclear if the difference in problem identification originates from this change. However, to explore the possible influence of their presence it seems that their presence cannot be simulated. Their relationship to the participant has to be genuine, and cannot be simulated. Similarly the data of the usability evaluation has to be actual data concerning the non-operative present, and the tasks being solved have to actually relate to the non-operative present. If this is not fulfilled, it seems that it affects the way the participant perceives the situation and the presence of the non-operative. This is also supported by literature observing that participants behave differently, when they know that their actions have real and actual impact on non-operatives, with whom they have a genuine relationship, whereas consequences are not considered when the non-operatives simulate a relationship, since the participant knows that the consequences are simulated too and thus not real.

5.4 Limitations and Further Work

The research of this thesis holds a number of limitations wit regards to the general validity of the results. Firstly, the choice of the theory of behaviour settings has introduced a perspective related to the power relations in social context to the research, and while this perspective has fit the area of research well, the interpretation of data might change with the application of a different theory with another perspective, yielding results regarding different aspects of social context.

Secondly, the participants involved in my experiments had very distinct profiles, (children and home healthcare workers) with the characteristics that these groups typically possess. Therefore the results may not be generalisable to other user groups with notably different characteristics than the user groups involved.

Thirdly, it has become clear in my research on the influence of non-operatives on usability evaluations, that to be able to observe their influence, it is important that the social context of the usability evaluation is genuine. It cannot be simulated or

recreated, as important aspects such as genuine relationship and consequences of actions have major impact on how non-operatives influence the participant with regards to verbalisation, collaboration and problem identification. Based on the experience gained from the research of this thesis, the impact of non-operatives on usability evaluations is a complex area of research, requiring great attention to details to produce valid results.

Finally, this thesis has focused mainly on gaining an understanding of the influence of social context on usability evaluations through a number of usability evaluations manipulating the social context. Therefore, due to time constraints, it was not within the scope of this thesis engineering e.g. a new usability evaluation method focusing on the incorporation of social context, which is a limitation.

In response to the limitations above, future research may include the development of a new usability evaluation method, based on the findings and experiences of this thesis. This new method should be aimed specifically at incorporating social context in the evaluation. Furthermore, a more thorough study of non-operatives and their impact on usability evaluations is needed and should be planned, taking the experience gained in this thesis into account.

References

Als, B. S., Jensen, J. J. & Skov, M. B. (2005) Exploring Verbalization and Collaboration of Constructive Interaction with Children. *Proceedings of the 10th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'05)*, 443-456, Berlin: Springer-Verlag.

Als, B. S., Jensen, J. J. & Skov, M. B. (2009) Composing Children Dyads in Constructive Interaction: A Comparison of Usability Testing Methods for Problem Identification. (Extended version of Als, B. S., Jensen, J. J. & Skov, M. B. (2005) Comparison of Think-Aloud and Constructive Interaction in Usability Testing with Children. *Proceedings of the 4th International Conference for Interaction Design and Children (IDC'05)*, 80-87, New York: ACM Press.)

Barker, R. G. (1978) *Ecological Psychology – Concepts and Methods for Studying the Environment of Human Behavior*. Stanford, CA, Stanford University Press.

Barker, Roger G. & Schoggen, P. (1978) Measures of Habitat and Behavior Output. In Barker, Roger G. and associates (Eds.) *Habitats, Environments, and Human Behaviour – Studies in Ecological Psychology and Eco-Behavioral Science from the Midwest Psychological Field Station, 1947-1972.* 229-244, San Francisco: Jossey-Bass Publishers.

Basili, V. R., Selby, R. W. & Hutchins, D. H. (1986) Experimentation in Software Engineering. *IEEE Transactions on Software Engineering, 12*(7), 733-743, IEEE Press.

Bekker, M., Beusmans, J., Keyson, D. & Lloyd, P. (2003) KidReporter: A User Requirements Gathering Technique for Designing with Children. *Interacting with Computers (IwC), 15*(2), 187-202, Elsevier.

Benford, S., Bederson, B. B., Åkesson, K.-P., Bayon, V., Druin, A., Hansson, P. et al. (2000) Designing Storytelling Technologies to Encouraging Collaboration between Young Children. *Proceedings of the 18th International Conference on Human Factors in Computing Systems (CHI'00)*, 556-563, New York: ACM Press.

Bers, M. U., Ackermann, E., Cassell, J., Donegan, B., Gonzalez-Heydrich, J., DeMaso, D. R. et al. (1998) Interactive Storytelling Environments: Coping with Cardiac Illness

at Boston's Children's Hospital. *Proceedings of the 16th International Conference on Human Factors in Computing Systems (CHI'98)*, 603-610, New York: ACM Press.

Bevan, N. & Macleod, M. (1994) Usability Measurement in Context. *Behaviour and Information Technology (BIT)*, 13(1-2), 132-145. Taylor & Francis Group.

Bias, R. G. & Mayhew, D. J. (Eds.) (1994) *Cost-Justifying Usability.* Academic Press

Blanchard, A. (2004) Virtual Behavior Settings: An Application of Behavior setting Theories to Virtual Communities. *Computer-Mediated Communication*, 9(2), IN: Indiana University.

Brooke, J. (1996) SUS – A Quick and Dirty Usability Scale. In Jordan, P. W., Thomas, B., Weerdmeester, B. A. & McClelland, I. L. (Eds.) *Usability Evaluation in Industry.* 189-194, London: Taylor & Francis Group.

Brown, P.J., Bovey, J.D. & Chen, X. (1997) Context-Aware Applications: From the Laboratory to the Marketplace. *IEEE Personal Communications*, 4(5), 58-64. IEEE Press.

Chen, G. & Kotz, D. (2000) *A Survey of Context-Aware Mobile Computing Research.* Technical report TR2000-381, Hannover, NH: Dartmouth College, Department of Computer Science.

Danesh, A., Inkpen, K., Lau, F., Shu, K. & Booth, K. (2001) GeneyTM: Designing a Collaborative Activity for the palmTM Handheld Computer. *Proceedings of the 19th International Conference on Human Factors in Computing Systems (CHI'01)*, 388-395, New York: ACM Press.

Dey, A.K. (1998) Context-Aware Computing: The CyberDesk Project. *AAAI 1998 Spring Symposium on Intelligent Environments*, Technical Report SS-98-02, 51-54.

Dey, A. K. & Abowd, G. D. (2000) Towards a Better Understanding of Context and Context-Awareness. In CHI2000 Workshop on What, Who, Where, When and How of Context-Awareness, New York: ACM Press.

Dix, A., Finlay, J., Abowd, G. D. & Beale, R. (2004) *Human Computer Interaction (3rd ed.).* Essex, England: Pearson Education.

Dumas, J. S. & Loring, B. A. (2008) *Moderating Usability Tests: Principles and Practice for Interacting.* USA: Morgan Kaufmann.

Dumas, J. S. & Redish, J. C. (1999). *A practical guide to usability testing. (2nd ed.).* Portland, OR: Intellect.

Ericsson, K. A. and Simon, H. A. (1993) *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: MIT Press.

Frohlich, D. & Kraut, R. (2003) The Social Context of Home Computing. In Harper, R. (Ed.) *Inside the Smart Home.* 127-162, London: Springer- Verlag

Galliers, R. D. (1992) Choosing Information Systems Research Approaches. In Galliers, R. D. (ed.) *Information Systems Research: Issues, Methods and Practical Guidelines.* 144-162, Boston: Blackwells Scientific Publications.

Gutwin, C. & Greenberg, S. (2000) The Mechanics of Collaboration: Developing Low Cost Usability Evaluation Methods for Shared Workspaces. *In IEEE 9th International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, 98-103, Gaithersburg, MD: IEEE Press

Hanna, L., Risden, K., and Alexander, K. J. (1997) Guidelines for Usability Testing with Children. *Interactions Magazine, 4*(5), 9-14, New York: ACM.

Heim, S. (2008) *The Resonant Interface – HCI Foundations for Interaction Design.* USA: Pearson-Addison Wesley.

Hillier, M. (2003) The Role of Cultural Context in Multilingual Website Usability. *Electronic Commerce Research and Application*, 2(1), 2-14, Elsevier.

Høegh, R. Th. & Jensen, J. J. (2008) A Case Study of Three Software Projects: Can Software Developers Anticipate the Usability Problems in their Software? *Behaviour & Information Technology (BIT)*, 27(4) 307-312, Taylor & Francis Group.

Höysniemi, J., Hämäläinen, P. & Turkki, L. (2003) Using Peer Tutoring in Evaluating the Usability of a Physically Interactive Computer Game with Children. In *Interacting with Computers (IwC)*, 15(2), 203-225, Elsevier.

Jensen, J. J. (2007) Evaluating in a Healthcare Setting: A Comparison between Concurrent and Retrospective Verbalisation. *Proceedings of the 12th International Conference on Human-Computer Interaction - Interaction Design and Usability*, 508-516, Berlin: Springer-Verlag.

Jensen, J. J. & Skov, M. B. (2009) A Classification of Research Methods and Purposes in Child-Computer Interaction. (Extended version of Jensen, J. J. & Skov, M. B. (2005) A Review of Research Methods in Children's Technology Design. *Proceedings of the 4th International Conference for Interaction Design and Children (IDC'04)*, 9-16, Boulder, CO).

Jones, M. & Marsden, G. (2006) *Mobile Interaction Design.* West Sussex, England: John Wiley and Sons.

Kahler, H. (2000) Constructive Interaction and Collaborative Work: Introducing a Method for Testing Collaborative Systems, *Interactions Magazine, 7*(3) , 27-34, New York: ACM.

Kjeldskov J. and Graham C. (2003) A Review of Mobile HCI Research Methods. *Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'03)*, 317-335, Berlin: Springer-Verlag.

Kjeldskov, J. & Skov, M. B. (2003) Creating Realistic Laboratory Settings: Comparative Studies of Three Think-Aloud Usability Evaluations of a Mobile System. *Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'03)*, 663-670, IOS Press.

Kjeldskov, J., Skov, M. B., Als, B. S. & Høegh, R. Th. (2004) Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. *Proceedings of the 6th International Conference on Human-Computer Interaction with Mobile Devices and services (MobileHCI'04)*, 61-73, Berlin: Springer-Verlag.

Kjeldskov, J. & Skov, M. B. (2007) Studying Usability In Sitro: Simulating Real World Phenomena in Controlled Environments. *International Journal of Human-Computer Interaction*, 22(1-2), 7-36, Taylor & Francis Group.

Kjeldskov, J., & Stage, J. (2003). The Process of Developing a Mobile Device for Communication in a Safety-Critical Domain. *Proceedings of the 9th IFIP TC13 International Conference on Human Computer Interaction (INTERACT'03),* 264–271, IOS Press.

Kjeldskov, J. & Stage, J. (2004) New Techniques for Usability Evaluation of Mobile Systems. *International Journal of Human-Computer Studies,* 60(4-5), 599-620, Taylor & Francis Group.

Leventhal, L. & Barnes, J. (2008) *Usability Engineering - Process, Products and Examples.* Upper Saddle River, NJ: Pearson-Prentice Hall.

Maguire, M. (2001) Context of Use within Usability Activities. *International journal of Human-Computer Studies,* 55(4), 453-483, Taylor & Francis Group.

Markopoulos, P. & Bekker, M. (2003) On the Assessment of Usability Testing Methods for Children. *Interacting with Computers (IwC)*, 15(2), 227-243, Elsevier.

Markopoulos, P., Read, J. C., MacFarlane, S. & Höysniemi, Johanna (2008) *Evaluating Children's Interactive Products: Principles and Practices for Interaction Designers.* Burlington, MA: Morgan Kaufmann.

Montemayor, J., Druin, A., Farber, A., Simms, S., Churaman, W. & D'Amour, A. (2002) Physical Programming: Designing Tools for Children to Create Physical Interactive Environments. *Proceedings of the 20th International Conference on Human Factors in Computing Systems (CHI'02)*, 299-306, New York: ACM.

Nielsen, J. (1993) *Usability Engineering.* San Francisco, Morgan Kaufmann

Nielsen, C. M., Overgaard, M., Pedersen, M. B., Stage, J. & Stenild, S. (2006) It's Worth the Hassle! The Added Value of Evaluating the Usability of Mobile Systems in the Field. *Proceedings of the 4th Nordic Conference on Human-Computer Interaction (NordiCHI'06): Changing Roles*, 272-280, New York: ACM.

Nivala, A.-M. & Sarjakoski, L. T. (2003) Need for Context-Aware Topographic Maps in Mobile Devices. *Proceedings of the 9th Scandinavian Research Conference on Geographical Information Science (ScanGIS'03)*, 15-29.

O'Malley, C. E., Draper, S. W. & Riley, M. S. (1984) Constructive Interaction: A Method for Studying Human-Computer-Human Interaction. *Proceedings of the 1st IFIP TC13 International Conference on Human Computer Interaction (INTERACT'84)*, 269-274, London: North-Holland.

Po, S., Howard, S., Vetere, F. & Skov, M. B. (2004) Heuristic Evaluation and Mobile Usability: Bridging the Realism Gap. *Proceedings of the 6th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'04)*, 61-73, Berlin: Springer-Verlag.

Preece, J., Rogers, Y. & Sharp, H. (2007) *Interaction Design - Beyond Human-Computer Interaction (2nd ed.).* West Sussex, England: John Wiley & Sons.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. & Carey, T. (1994) *Human-Computer Interaction.* Addison-Wesley.

Rauterberg M. (2003) *Cost Justifying Usability - A State of the Art Overview.* Technical Report TUE-ID-004-03, The Netherlands, Technical University of Eindhoven.

Rubin, J. (1994) *Handbook of Usability Testing – How to Plan, Design, and Conduct Effective Tests.* USA: John Wiley and Sons.

Ryan, N., Pascoe, J. & Morse, D. (1997) Enhanced Reality Fieldwork: the Context-Aware Archaeological Assistant. In Gaffney, V., van Leusen, M. & Exxon, S. (Eds.) *Computer Applications in Archaeology*.

Schilit, B. & Theimer, M. (1994) Disseminating Active Map Information to Mobile Hosts. *IEEE Network*, 8(5), 22-32. IEEE Press.

Schneiderman, B. (1998) *Designing the User Interface – Strategies for Effective Human-Computer Interaction (3rd ed.).* Addison-Wesley.

Schoggen, P. (1989) *Behavior Settings: A Revision and Extension of Roger G. Barker's 'Ecological Psychology'*. Stanford, CA: Stanford University Press.

Scott Morton, M. (1985) The State of the Art of Research. In F. W. McFarlan (Ed.) *The Information Systems Research Challenge*. 13-41, Boston: Harvard Business School Press.

van den Haak, M. J. & de Jong, M. D. T. (2005) Analyzing the Interaction between Facilitator and Participants in Two Variants of the Think-Aloud Method. *Proceedings of the IEEE International Professional Communication Conference*, 323-327, IEEE Press.

van Kesteren, I. E. H., Bekker, M. M., Vermeeren, A. P. O. S., and Lloyd, P. A. (2003) Assessing Usability Evaluation Methods on Their Effectiveness to Elicit Verbal Comments from Children Subjects. *Proceedings of the 2nd International Conference for Interaction Design and Children (IDC'03)*, 41-49, New York: ACM Press.

Wicker, A. W. (1979) *An Introduction to Ecological Psychology*. Monterey, Canada, Brooks Cole Publishing.

Wicker, A. W. (1987) Behavior Settings Reconsidered: Temporal Stages, Resources, Internal Dynamics, Context. In *Handbook of Environmental Psychology, Vol. 1.*, New York: John Wiley & Sons.

Wicker, A. W. (1992) Making Sense of Environments. *Person-Environment Psychology: Models and Perspectives*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Wildman, D. (1995) Getting the Most from Paired-User Testing. *Interactions Magazine*, *2*(3), 21–27, New York: ACM.

Wilson, C. & Blostein, J. (1998) Pros and Cons of Co-Participation in Usability Studies. *Usability Interface*, 4(4).

Wixon, D. (2003**)** Evaluating Usability Methods: Why the Current Literature Fails the Practitioner. *Interactions Magazine, 10*(4), 28–34, New York: ACM.

Wynekoop, J. L. and Congor, A. A. (1990) A Review of Computer Aided Software Engineering Research Methods. *Proceedings of the IFIP TC8 WG8.2 Working Conference on the Information Systems Research Arena of the 90's*, 129-154.

Appendix A Paper Contributions

1. Als, B. S., Jensen, J. J. & Skov, M. B. (2005) Exploring Verbalization and Collaboration of Constructive Interaction with Children. *Proceedings of the 10th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'05)*, 443-456, Berlin: Springer-Verlag.

2. Høegh, R. Th. & Jensen, J. J. (2008) A Case Study of Three Software Projects: Can Software Developers Anticipate the Usability Problems in their Software? *Behaviour & Information Technology (BIT)*, 27(4) 307-312, Taylor & Francis Group.

3. Als, B. S., Jensen, J. J. & Skov, M. B. (2009) Composing Children Dyads in Constructive Interaction: A Comparison of Usability Testing Methods for Problem Identification. (Extended version of Als, B. S., Jensen, J. J. & Skov, M. B. (2005) Comparison of Think-Aloud and Constructive Interaction in Usability Testing with Children. *Proceedings of the 4th International Conference for Interaction Design and Children (IDC'05)*, 80-87, New York: ACM Press.)

4. Jensen, J. J. (2007) Evaluating in a Healthcare Setting: A Comparison between Concurrent and Retrospective Verbalisation. *Proceedings of the 12th International Conference on Human-Computer Interaction - Interaction Design and Usability*, 508-516, Berlin: Springer-Verlag.

5. Jensen, J. J. & Skov, M. B. (2009) A Classification of Research Methods and Purposes in Child-Computer Interaction. (Extended version of Jensen, J. J. & Skov, M. B. (2005) A Review of Research Methods in Children's Technology Design. *Proceedings of the 4th International Conference for Interaction Design and Children (IDC'04)*, 9-16, Boulder, CO).

Appendix B Re-categorisation of papers

The 87 papers that were categorised as having the purpose of evaluation in Jensen and Skov (2009), have for the benefit of this summary been re-categorised according to the framework introduced in chapter 2. This appendix will explain the background and procedure of this re-categorisation along with any problems related to the procedure, the results of the re-categorisation and supporting examples.

In Jensen and Skov (2009), 132 papers on child computer interaction were categorised according to the framework of research method and research purpose by Wynekoop and Conger (1990). Out of these 132 papers, 87 papers were categorised as having the purpose of evaluation. The methods applied were laboratory experiment and field study. The papers were divided roughly evenly between the two methods, with some papers falling into both categories. As the subject of this thesis is social context in usability evaluations, I chose to re-categorise these papers according to the framework previously introduced to extract current practice with regards to social context in usability evaluations.

The re-categorisation was done by the author of this thesis and started out with a trial categorisation of 10 randomly selected papers. Over a period of three weeks all 87 papers were then read and categorised. A number of uncertainties arose and were addressed through discussion with the co-author of Jensen and Skov (2009). These uncertainties arose mainly because many papers were less than clear on the level of involvement of the people in the evaluation as well as the context of the evaluation and those papers required a bit of interpretation on the author's part. Upon reaching agreement on the uncertainties mentioned, all papers were re-read to verify the initial categorisation. The categorisation of a few papers changed based on the discussion and the categorisation was finalised. A list of the re-categorised papers can be seen in section B.1 Note that the numbers of the papers are the original numbers of Jensen & Skov (2009). These were kept to allow for reference to the original review, rather than renumber the papers. Thus, the numbers are not sequential, and in the paper list the 'empty' numbers represent the papers that were not re-categorised.

In some cases a paper would fall in more than one category such as papers reporting problems connected to the system being evaluated, but also reporting on issues discovered in connection to the method applied, or papers applying constructive interaction in a field study, thus having both operatives and non-operatives involved in the evaluation.

The categorisation resulted in the distribution of papers illustrated in table 5 (see next page). Out of the 87 papers doing evaluation, only 64 of them reported on some form of social context. Thus 23 papers fell outside the framework. These were typically

papers reporting on single testers, with no test leader present or no empirical findings regarding the test leader. Out of the 64 papers in the framework, nearly all (61) included some form of operatives in their evaluation (typically children evaluating in pairs or groups), whereas only 10 papers included non-operatives, which supports earlier claims that non-operatives is a subject dealt with by very few researchers. Furthermore, all of these 10 papers include non-operatives solely as a remark and not as a focus of their research. The division between product and process is slightly more evened out as 34 papers report on process while 58 papers report on product. The papers in the process category typically report on findings concerning a method employed or regarding verbalisation or collaboration between participants and/or test leader, while the papers in the product category report on various findings regarding a specific application or system being evaluated. 33 papers report on the process of the usability evaluation in connection to the operatives involved, 55 papers report on the product of the usability evaluation in connection to the operatives involved, only 5 papers report on the process of the usability evaluation in connection to the non-operatives involved and only 8 papers report on the product of the usability evaluation in connection to the non-operatives involved.

		Social Context (64)	
		Operatives (61)	**Non-operatives** (10)
Usability Evaluation (64)	**Process** (34)	33 papers: 4, 5, 6, 9, 10, 19, 24, 25, 28, 29, 31, 44, 48, 50, 51, 56, 57, 69, 74, 78, 89, 90, 91, 92, 96, 97, 109, 112, 117, 121, 127, 130, 131	5 papers: 10, 15, 48, 117, 127
	Product (58)	55 papers: 3, 5, 6, 11, 12, 17, 18, 19, 21, 24, 25, 28, 29, 31, 37, 38, 39, 42, 44, 47, 48, 50, 51, 53, 54, 57, 73, 78, 80, 83, 88, 89, 90, 91, 92, 96, 97, 98, 99, 100, 106, 107, 109, 110, 112, 114, 117, 119, 120, 121, 122, 127, 128, 130, 131	8 papers: 15, 16, 39, 48, 52, 73, 119, 127
Out of category (23)		23 papers: 7, 14, 22, 23, 30, 59, 62, 64, 66, 67, 75, 76, 82, 85, 94, 95, 101, 104, 105, 113, 123, 124, 132	

Table 5: The distribution of the 87 papers doing evaluation, between the categories Operatives/Non-operatives and Product/Process. The numbers in parentheses by each category are the number of papers associated with that category.

B.1 Reviewed Research Papers

1.

2.

3. Als, B. S., Jensen, J. J. & Skov, M. B. (2005) Comparison of Think-Aloud and Constructive Interaction in Usability Testing with Children. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, 9-16, New York: ACM Press.

4. Als, B. S., Jensen, J. J. & Skov, M. B. (2005) Exploring Verbalization and Collaboration of Constructive Interaction with Children. *Proceedings of the 10th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'05)*, 443-456, IOS Press.

5. Antle, A. (2003) Case Study: The Design of CBC4Kids' Storybuilder. *Proceedings of the 2nd International Conference on Interaction Design and Children (IDC'03)*, 59-68, New York: ACM Press.

6. Antle, A: (2004) Supporting Children's Emotional Expression and Exploration in Online Environments. *Proceedings of the 3rd International Conference on Interaction Design and Children (IDC'04)*, 97-104, New York: ACM Press.

7. Baauw, E., Bekker, M. M. & Barendregt, W. (2005) A Structured Expert Evaluation Method for the Evaluation of Children's Computer Games. *Proceedings of the 10th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'05)*, 457-469, Berlin: Springer-Verlag.

8.

9. Bekker, M, Beusmans, J., Keyson, D. & Lloyd, P. (2002) KidReporter: A Method for Engaging Children in Making a Newspaper to Gather User Requirements. *Proceedings of the 1st International Conference on Interaction Design and Children (IDC'02).*

10. Bekker, M, Beusmans, J., Keyson, D. & Lloyd, P. (2003) KidReporter: A User Requirements Gathering Technique for Designing with Children. *Interacting with Computers (IwC)*, 15(2), 187-202, Elsevier.

11. Benford, S., Bederson, B. B., Åkesson, K-P, Bayon, V., Druin, A., Hansson, P. et al. (2000) Designing Storytelling Technologies to Encourage Collaboration between Young Children. *Proceedings of the 18th International Conference on Human Factors in Computing Systems (CHI'00)*, 556-563, New York: ACM Press.

12. Benford, S., Rowland, D., Flintham, M., Drozd, A., Hull, R., Reid, J. et al. (2005) Life on the Edge: Supporting Collaboration in Location-Based Experiences. *Proceedings of the 23rd International Conference on Human Factors in Computing Systems (CHI'05)*, 721-730, New York: ACM Press.

13.

14. Bernard, M. L., Chaparro, B. S., Mills, M. M. & Halcomb, C. G. (2002) Examining Children's Reading Performance and Preference for Different Computer-Displayed Text. *Behaviour and Information Technology (BIT)*, 21(2), 87-96, Taylor & Francis Group.

15. Bers, M. U., Ackermann, E., Cassell, J., Donegan, B., Gonzalez-Heydrich, J., DeMaso, D. R. et al. (1998) Interactive Storytelling Environments: Coping with Cardiac Illness at Boston's Children's Hospital. *Proceedings of the 16th International Conference on Human Factors in Computing Systems (CHI'98)*, 603-610, New York: ACM Press.

16. Bers, M. U., Gonzalez-Heydrich, J. & DeMaso, D. R. (2001) Identity Construction Environments: Supporting a Virtual Therapeutic Community of Pediatric Patients Undergoing Dialysis. *Proceedings of the 19th International Conference on Human Factors in Computing Systems (CHI'01)*, 380-387, New York: ACM Press.

17. Borovoy, R., Silverman, B., Gorton, T., Klann, J., Notowidigdo, M., Knep, B. et al. (2001) Folk Computing: Revisiting Oral Tradition as a Scaffold for Co-Present Communities. *Proceedings of the 19th International Conference on Human Factors in Computing Systems (CHI'01)*, 466-473, New York: ACM Press.

18. Bouvin, N. O., Brodersen, C., Hansen, F. A., Iversen, O. S. & Nørregaard, P. (2005) Tools of Contextualization: Extending the Classroom to the Field. In *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, 24-31, New York: ACM Press.

19. Brederode, B., Markopoulos, P., Gielen, M., Vermeeren, A. & de Ridder, H. (2005) pOwerball: The Design of a Novel Mixed-Reality Game for Children with Mixed Abilities. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, 32-39, New York: ACM Press.

20.

21. Cassell, J. & Ryokai, K. (2001) Making Space for Voice: Technologies to Support Children's Fantasy and Storytelling. *Personal and Ubiquitous Computing (PUC)*, 5(3), 169-190, New York: ACM Press.

22. Chen, C-H., Wu, F-G., Rau, P-L. P. & Hung, Y-H. (2004) Preferences of young children regarding interface layouts in child community web sites. *Interacting with Computers (IwC)*, 16(2), 311-330, Elsevier.

23. Chiasson, S. & Gutwin, C. (2005) Testing the Media Equation with Children. *Proceedings of the 23rd International Conference on Human Factors in Computing Systems (CHI'05)*, 829-838, New York: ACM Press.

24. Cockburn, A. & Greenberg, S. (1998) The Design and Evolution of TurboTurtle, a Collaborative Microworld for Exploring Newtonian Physics. *International Journal of Human-Computer Studies (IJHCS)*, 48(6), 777-801, New York: ACM Press.

25. Danesh, A., Inkpen, K., Lau, F., Shu, K. & Booth, K. (2001) GeneyTM: Designing a Collaborative Activity for the palmTM Handheld Computer. *Proceedings of the 19th International Conference on Human Factors in Computing Systems (CHI'01)*, 388-395, New York: ACM Press.

26.

27.

28. Decortis, F., Rizzo, A. & Saudelli, B. (2003) Mediating Effects of Active and Distributed Instruments on Narrative Activities. *Interacting with Computers (IwC)*, 15(6), 801-830, Elsevier.

29. Dindler, C., Eriksson, E., Iversen, O. S., Lykke-Olesen, A. & Ludvigsen, M. (2005) Mission from Mars – A Method for Exploring User Requirements for Children in a Narrative Space. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, 40-47, New York: ACM Press.

30. Donker, A. & Reitsma, P. (2004) Usability testing with young children. In *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'04)*, 43-48, New York: ACM Press.

31. Druin, A., Stewart, J., Proft, D., Bederson, B. & Hollan, J. (1997) KidPad: A Design Collaboration between Children, Technologists, and Educators. *Proceedings of the 15th International Conference on Human Factors in Computing Systems (CHI'97)*, 463-470, New York: ACM Press.

32.

33.

34.

35.

36.

37. Ellis, J. B. & Bruckman, A. S. (2001) Designing Palaver Tree Online: Supporting Social Roles in a Community of Oral History. *Proceedings of the 19th International Conference on Human Factors in Computing Systems (CHI'01)*, 474-481, New York: ACM Press.

38. Fails, J. A., Druin, A., Guha, M. L., Chipman, G., Simms, S. & Churaman, W. (2005) Child's Play: A Comparison of Desktop and Physical Interactive Environments. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, 48-55, New York: ACM Press.

39. Fels, D. I., Waalen, J. K., Zhai, S. & Weiss, P. T. (2001) Telepresence Under Exceptional Circumstances: Enriching the Connection to School for Sick Children. *Proceedings of the 8th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'01)*, 617-624, IOS Press.

40.

41.

42. Frei, P., Su, V., Mikhak, B. & Ishii, H. (2000) curlybot: Designing a New Class of Computational Toys. *Proceedings of the 18th International Conference on Human Factors in Computing Systems (CHI'00)*, 129-136, New York: ACM Press.

43.

44. Gibson, L., Newall, F. & Gregor, P. (2003) Developing a Web Authoring Tool that Promotes Accessibility in Children's Design. *Proceedings of the 2nd International Conference on Interaction Design and Children (IDC'03)*, 23-30, New York: ACM Press.

45.

46.

47. Gweon, G., Ngai, J. & Rangos, J. (2005) Exposing Middle School Girls to Programming via Creative Tools. *Proceedings of the 8th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'05)*, 431-442, Berlin: Springer-Verlag.

48. Hall, T. & Bannon, L. (2005) Designing Ubiquitous Computing to Enhance Children's Interaction in Museums. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, 62-69, New York: ACM Press.

49.

50. Hanna, L., Neapolitan, D. & Risden, K. (2004) Evaluating Computer Game Concepts with Children. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'04)*, 49 - 56, New York: ACM Press.

51. Henderson, V., Lee, S., Brashear, H., Hamilton, H., Starner, T. & Hamilton, S. (2005) Development of an American Sign Language Game for Deaf Children. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, 70-79, New York: ACM Press.

52. Hornof, A. J. & Cavender, A. (2005) EyeDraw: Enabling Children with Severe Motor Impairments to Draw with Their Eyes. *Proceedings of the 23rd International Conference on Human Factors in Computing Systems (CHI'05)*, 161 - 170, New York: ACM Press.

53. Hourcade, J. P., Bederson, B. B., Druin, A., Rose, A., Farber, A. & Takayama, Y. (2002) The International Children's Digital Library: Viewing Digital Books Online. *Proceedings of the 1st International Conference on Interaction Design and Children (IDC'02)*

54. Hourcade, J. P., Bederson, B. B., Druin, A., Rose, A., Farber, A. & Takayama, Y. (2003) The International Children's Digital Library: Viewing Digital Books Online. *Interacting with Computers (IwC)*, 15(2), 151-167, Elsevier.

55.

56. Höysniemi, J., Hämäläinen, P. & Turkki, L. (2002) Using Peer Tutoring in Evaluating Usability of Physically Interactive Computer Game with Children. *Proceedings of the 1st International Conference on Interaction Design and Children (IDC'02)*

57. Höysniemi, J., Hämäläinen, P. & Turkki, L. (2003) Using Peer Tutoring in Evaluating the Usability of a Physically Interactive Computer Game with Children. *Interacting with Computers (IwC)*, 15(2), 203-225, Elsevier.

58.

59. Inkpen, K. M. (2001) Drag-and-Drop versus Point-and-Click Mouse Interaction Styles for Children. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 8 (1), 1-33, New York: ACM Press.

60.

61.

62. Jacko, J. A. (1996) The Identifiability of Auditory Icons for Use in Educational Software for Children. *Interacting with Computers (IwC)*, 8(2), 121-133, Elsevier.

63.

64. Johnson, M. P., Wilson, A., Blumberg, B., Kline, C. & Bobick, A. (1999) Sympathetic Interfaces: Using a Plush Toy to Direct Synthetic Characters. *Proceedings of the 17th International Conference on Human Factors in Computing Systems (CHI'99)*, 152-158, New York: ACM Press.

65.

66. Jung, Y., Persson, P. & Blom, J. (2005) DeDe: Design and Evaluation of a Context-Enhanced Mobile Messaging System. *Proceedings of the 23rd International Conference on Human Factors in Computing Systems (CHI'05)*, 351-360, New York: ACM Press.

67. Kaplan, N. & Chisik, Y. (2005) Reading Alone Together: Creating Sociable Digital Library Books. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, 88-94, New York: ACM Press.

68.

69. van Kesteren, I. E. H., Bekker, M. M., Vermeeren, A. P. O. S. & Lloyd, P. A. (2003) Assessing Usability Evaluation Methods on Their Effectiveness to Elicit Verbal Comments from Children Subjects. *Proceedings of the 2nd International Conference on Interaction Design and Children (IDC'03)*, 41-49, New York: ACM Press.

70.

71.

72.

73. Labrune, J.-P. & Mackay, W. (2005) Tangicam: Exploring observation tools for children. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, 95-102, New York: ACM Press.

74. Lamberty, K. K. & Kolodner, J. L. (2005) Camera Talk: Making the Camera a Partial Participant. *Proceedings of the 23rd International Conference on Human Factors in Computing Systems (CHI'05)*, 839-848, New York: ACM Press.

75. Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A. & Bhogal, R. S. (1997) The Persona Effect: Affective Impact of Animated Pedagogical Agents. In *Proceedings of the 15th International Conference on Human Factors in Computing Systems (CHI'97)*, 359-366, New York: ACM Press.

76. Lewis, C., Brand, C., Cherry, G. & Rader, C. (1998) Adapting User Interface Design Methods to the Design of Educational Activities. *Proceedings of the 16th International Conference on Human Factors in Computing Systems (CHI'98)*, 619-626, New York: ACM Press.

77.

78. Loh, B., Radinsky, J., Russell, E., Gomez, L. M., Reiser, B. J. & Edelson, D. C. (1998) The Progress Portfolio: Designing Reflective Tools for a Classroom Context. *Proceedings of the 16th International Conference on Human Factors in Computing Systems (CHI'98)*, 627-634, New York: ACM Press.

79.

80. Lumbreras, M. & Sánchez, J. (1999) Interactive 3D Sound Hyperstories for Blind Children. *Proceedings of the 17th International Conference on Human Factors in Computing Systems (CHI'99)*, 318-325, New York: ACM Press.

81.

82. Markopoulos, P. & Bekker, M. (2002) How to Compare Usability Testing Methods with Children Participants. *Proceedings of the 1st International Conference on Interaction Design and Children (IDC'02)*

83. Markopoulos, P. & Bekker, M. (2003) On the Assessment of Usability Testing Methods for Children. *Interacting with Computers (IwC)*, 15(2), 227-243, Elsevier.

84.

85. McElligott, J. & van Leeuwen, L. (2004) Designing Sound Tools and Toys for Blind and Visually Impaired Children. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'04)*, 65-72, New York: ACM Press.

86.

87.

88. Moher, T., Johnson, A., Ohlsson, S. & Gillingham, M. (1999) Bridging Strategies for VR-based Learning. *Proceedings of the 17th International Conference on Human Factors in Computing Systems (CHI'99)*, 536-543, New York: ACM Press.

89. Montemayor, J., Druin, A., Farber, A., Simms, S., Churaman, W. & D'Amour, A. (2002) Physical Programming: Designing Tools for Children to Create Physical Interactive Environments. *Proceedings of the 20th International Conference on Human Factors in Computing Systems (CHI'02)*, 299-306, New York: ACM Press.

90. Mäkelä, A., Giller, V., Tscheligi, M. & Sefelin, R. (2000) Joking, Storytelling, Artsharing, Expressing Affection: A Field Trial of how Children and Their Social Network Communicate with Digital Images in Leisure Time. *Proceedings of the 18th International Conference on Human Factors in Computing Systems (CHI'00)*, 548-555, New York: ACM Press.

91. Oosterholt, R., Kusano, M. & de Vries, G. (1996) Interaction Design and Human Factors Support in the Development of a Personal Communicator for Children. *Proceedings of the 14th International Conference on Human Factors in Computing Systems (CHI'96)*, 450-457, New York: ACM Press.

92. Ovaska, S., Hietala, P. & Kangassalo, M. (2003) Electronic Whiteboard in Kindergarten: Opportunities and Requirements. *Proceedings of the 2nd International Conference on Interaction Design and Children (IDC'03)*, 15-22, New York: ACM Press.

93.

94. Paiva, A., Andersson, G., Höök, K., Mourão, D., Costa, M. & Martinho, C. (2002) SenToy in FantasyA: Designing an Affective Sympathetic Interface to a Computer Game. *Personal and Ubiquitous Computing (PUC)*, 6(5-6), 378-389, New York: ACM Press.

95. Parés, N., Carreras, A., Durany, J., Ferrer, J., Freixa, P., Gómez, D. et al. (2005) Promotion of Creative Activity in Children with Severe Autism through Visuals. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, 110-116, New York: ACM Press.

96. Price, S., Rogers, Y., Scaife, M., Stanton, D. & Neale, H. (2002) Using 'Tangibles' to Promote Novel Forms of Playful Learning. *Proceedings of the 1st International Conference on Interaction Design and Children (IDC'02)*

97. Price, S., Rogers, Y., Scaife, M., Stanton, D. & Neale, H. (2003) Using 'Tangibles' to Promote Novel Forms of Playful Learning. *Interacting with Computers (IwC)*, 15(2), 169-185, Elsevier.

98. Rader, C., Brand, C. & Lewis, C. (1997) Degrees of Comprehension: Children's Understanding of a Visual Programming Environment. *Proceedings of the 15th International Conference on Human Factors in Computing Systems (CHI'97)*, 351-358, New York: ACM Press.

99. Raffle, H. S., Parkes, A. J. & Ishii, H. (2004) Topobo: A Constructive Assembly System with Kinetic Memory. *Proceedings of the 22nd International Conference on Human Factors in Computing Systems (CHI'04)*, 647-654, New York: ACM Press.

100. Randell, C., Price, S., Rogers, Y., Harris, E. & Fitzpatrick, G. (2004) The Ambient Horn: Designing a Novel Audio-Based Learning Experience. *Personal and Ubiquitous Computing (PUC)*, 8(3-4), 177-183, Berlin: Springer-Verlag.

101. Read, J. C., MacFarlane, S. & Casey, C. (2003) What's Going On? Discovering what Children Understand about Handwriting Recognition Interfaces. *Proceedings of the 2nd International Conference on Interaction Design and Children (IDC'03)*, 135-140, New York: ACM Press.

102.

103.

104. Resnick, M., Martin, F., Berg, R., Borovoy, R., Colella, V., Kramer, K. et al. (1998) Digital Manipulatives: New Toys to Think With. *Proceedings of the 16th International Conference on Human Factors in Computing Systems (CHI'98)*, 281-287, New York: ACM Press.

105. Risden, K., Czerwinski, M., Worley, S., Hamilton, L., Kubiniec, J., Hoffman, H. et al. (1998) Interactive Advertising: Patterns of Use and Effectiveness. *Proceedings of the 16th International Conference on Human Factors in Computing Systems (CHI'98)*, 219-224, New York: ACM Press.

106. Robertson, J. & Good, J. (2003) Ghostwriter: A Narrative Virtual Environment for Children. *Proceedings of the 2nd International Conference on Interaction Design and Children (IDC'03)*, 85-91, New York: ACM Press.

107. Robertson, J. & Good, J. (2004) Children's Narrative Development through Computer Game Authoring. *Proceedings of the 3rd International Conference on Interaction Design and Children (IDC'04)*, 57-64, New York: ACM Press.

108.

109. Rogers, Y., Price, S., Fitzpatrick, G., Fleck, R., Harris, E., Smith, H. et al. (2004) Ambient Wood: Designing New Forms of Digital Augmentation for Learning

Outdoors. *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'04)*, 3-10, New York: ACM Press.

110. Ryokai, K., Marti, S. & Ishii, H. (2004) I/O Brush: Drawing with Everyday Objects as Ink. *Proceedings of the 22nd International Conference on Human Factors in Computing Systems (CHI'04)*, 303-310, New York: ACM Press.

111.

112. Scaife, M., Rogers, Y., Aldrich, F. & Davies, M. (1997) Designing For or Designing with? Informant Design for Interactive Learning Environments. *Proceedings of the 15th International Conference on Human Factors in Computing Systems (CHI'97)*, 343-350, New York: ACM Press.

113. Scaife, M. & Rogers, Y. (2001) Informing the Design of a Virtual Environment to Support Learning in Children. *International Journal of Human-Computer Studies (IJHCS)*, 55(2), 115-143, New York: ACM Press.

114. Sharples, M., Corlett, D. & Westmancott, O. (2002) The Design and Implementation of a Mobile Learning Resource. *Personal and Ubiquitous Computing (PUC)*, 6(3), 220-234, New York: ACM Press.

115.

116.

117. Sluis, R. J. W., Weevers, I., van Schijndel, C. H. G. J., Kolos-Mazuryk, L., Fitrianie, S. & Martens, J. B. O. S. (2004) Read-It: Five-to-Seven-Year-Old Children Learn to Read in a Tabletop Environment. *Proceedings of the 3rd International Conference on Interaction Design and Children (IDC'04)*, 73 - 80, New York: ACM Press.

118.

119. Stanton, D., Bayon, V., Neale, H., Ghali, A., Benford, S., Cobb, S. et al. (2001) Classroom Collaboration in the Design Tangible Interfaces for Storytelling. *Proceedings of the 19th International Conference on Human Factors in Computing Systems (CHI'01)*, 482-489, New York: ACM Press.

120. Stewart, J., Bederson, B. B. & Druin, A. (1999) Single Display Groupware: A Model for Co-Present Collaboration. *Proceedings of the 17th International Conference on Human Factors in Computing Systems (CHI'99)*, 286-293, New York: ACM Press.

121. Stringer, M., Toye, E. F., Rode, J. A. & Blackwell, A. F. (2004) Teaching Rhetorical Skills with a Tangible User Interface. *Proceedings of the 3rd International Conference on Interaction Design and Children (IDC'04)*, 11-18, New York: ACM Press.

122. Strommen, E. F., Revelle, G. L., Medoff, L. M. & Razavi, S. (1996) Slow and Steady Wins the Race? Three-Year-Old Children and Pointing Device Use. *Behaviour and Information Technology (BIT)*, 15(1), 57-64, Taylor & Francis Group.

123. Strommen, E. (1998) When the Interface is a Talking Dinosaur: Learning Across Media with ActiMates Barney. *Proceedings of the 16th International Conference on Human Factors in Computing Systems (CHI'98)*, 288-295, New York: ACM Press.

124. Strommen, E. & Alexander, K. (1999) Emotional Interfaces for Interactive Aardvarks: Designing Affect into Social Interfaces for Children. *Proceedings of the 17th International Conference on Human Factors in Computing Systems (CHI'99)*, 528-535, New York: ACM Press.

125.

126.

127. Weiss, P. L., Whiteley, C. P., Treviranus, J. & Fels, D. I. (2001) PEBBLES: A Personal Technology for Meeting Educational, Social and Emotional Needs of Hospitalised Children. *Personal and Ubiquitous Computing (PUC)*, 5(3), 157-168, New York: ACM Press.

128. Williams, M., Jones, O. & Fleuriot, C. (2003) Wearable Computing and the Geographies of Urban Childhood - Working with Children to Explore the Potential of new Teechnology. *Proceedings of the 2nd International Conference on Interaction Design and Children (IDC'03)*, 111-118, New York: ACM Press.

129.

130. Wyeth, P. & Wyeth, G. (2001) Electronic Blocks: Tangible Programming Elements for Preschoolers. *Proceedings of the 8th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'01)*, 496-503, IOS Press.

131. Wyeth, P. & Purchase, H. C. (2003) Using Developmental Theories to Inform the Design of Technology for Children. *Proceedings of the 2nd International Conference on Interaction Design and Children (IDC'03)*, 93-100, New York: ACM Press.

132. Zuckerman, O., Arida, S. & Resnick, M. (2005) Extending Tangible Interfaces for Education: Digital Montessori-Inspired Manipulatives. *Proceedings of the 23rd International Conference on Human Factors in Computing Systems (CHI'05)*, 859-868, New York: ACM Press.

Exploring Verbalization and Collaboration of Constructive Interaction with Children

Benedikte S. Als, Janne J. Jensen, Mikael B. Skov

Department of Computer Science, Aalborg University
Fredrik Bajers Vej 7, DK-9220 Aalborg East, Denmark
{als, missj, dubois}@cs.aau.dk

Abstract. Constructive interaction provides natural thinking-aloud as test subjects collaborate in pairs to solve tasks. Since children may face difficulties in following instructions for a standard think-aloud test, constructive interaction has been suggested as evaluation method when usability testing with children. However, the relationship between think-aloud and constructive interaction is still poorly understood. We present an experiment that compares think-aloud and constructive interaction. The experiment involves 60 children with three setups where children apply think-aloud or constructive interaction in acquainted and non-acquainted pairs. Our results show that the pairing of children had impact on how the children collaborated in pairs and how they would afterward assess the testing sessions. In some cases, we found that acquainted dyads would perform well as they would more naturally interact and collaborate while in other cases they would have problems in controlling the evaluations.

1 Introduction

Children have been characterized as not just short adults, but as independent individuals with their own strong opinions, needs, likes, and dislikes, and they should be treated as such. The design and evaluation of children's technologies have received increased attention during the last several years [7, 8]. Druin [9] provides a classification of involvement where children play the roles of users, testers, informants, or design partners. The four roles encompass different levels of engagement and impose different opportunities and limitations. All roles involve different kinds of usability tests where children participate as subjects, for example user [29], tester [19], informant [10], and design partner [6].

Some research studies have started to investigate the roles of children in usability tests, cf. [18, 21]. Nielsen [26] suggests that evaluators should use a variation of think-aloud called constructive interaction [16, 23] (also known as co-discovery learning), since it may be difficult to get children to follow the instructions for a standard thinking-aloud test. Constructive interaction involves two test subjects collaborating in trying to solve tasks while using a computer system [27]. Even though constructive interaction with children seems appropriate, the relationship between think-aloud and constructive interaction in usability testing with children is poorly under-

stood. A number of questions still need to be addressed and answered: 1) How do children think-aloud and collaborate in constructive interaction 2) How should pairs of children be configured in constructive interaction? 3) How do children perceive the testing situation during constructive interaction?

In this paper, we investigate and address the above stated questions by looking at how children perform and behave in constructive interaction during usability testing. Our particular focus is on how the children behave and perceive a testing situation when involved a traditional think-aloud test compared to constructive interaction tests. First, we present an experimental design involving 60 children participating in two different configurations of constructive interaction and a traditional think-aloud. Secondly, we present results from the evaluations by illustrating how the children applied the think-aloud protocol and collaborated and further how they perceived the situation. Finally, we outline three lessons on involving children in usability testing.

2 Constructive Interaction in Usability Testing with Children

Nielsen [26] claims that constructive interaction is preferable over think-aloud when conducting usability evaluations with children. Where children face difficulties in following the instructions for a think-aloud test, constructive interaction comes closer to their natural behaviour, since the children work in pairs and collaborate in solving the tasks. Due to the fact that different the children's ability to verbalize their thoughts and feelings during a test, Hanna et al. [13] propose some adjusted guidelines where they reflect upon common target age ranges. Jensen and Skov [15] found that 67% of the research on interaction design and children applied some sort of systematic field or laboratory evaluations. Furthermore, some studies have explored different methods for conducting usability evaluations with children; one studied the effectiveness of co-operative evaluations (think-aloud) and co-discovery evaluations (constructive interaction) [1, 21], where another studied different method's effectiveness to elicit verbal comments from children [18]. The first compared the difference in total number of identifies usability problems identified by four subjects or four pairs, and found only negligible differences between the two methods.

Miyake [23] states that constructive interaction inherently integrates a number of opportunities and limitations. An advantage is that the test subjects naturally use think-aloud in their collaboration, one of the disadvantages is that the might aim for different strategies for learning and using computers. Furthermore, since constructive interaction requires twice as many test subjects as think-aloud, in order to conduct the same number of usability sessions, it is typically more expensive [26]. Configuring pairs for construction interaction includes two important steps [16]. First, test subjects must be selected and acquired for the usability test [27]. Secondly, usability evaluators are further faced with challenges of pairing subjects when adapting constructive interaction as evaluation technique. A number of challenges seem to influence the configuration of subjects in constructive interaction.

First, one challenge concerns the level of expertise. The level of expertise is important, as argued by O'Malley et al. [27], since the test subjects' knowledge of specific work tasks is quite often corresponding to their level of expertise. Nielsen [26]

recommends that the test subjects have the same level of experience, whereas having one of the test subjects enabled to guide the interaction, is an argument used by Kahler [16] when stating advantages by pairing test subjects with different levels of experience. Usually children do not posses expertise of work that might influence the outcome of the usability test, which makes the issue of expertise subtler when working with children. Most studies involving children do not explicitly consider the level of expertise [19, 25], one of the exceptions is a study where the participating children are profiled according to their scripting level [28]. Where age does not seem to matter when testing with adults, it has a more eloquent impact when conducting tests with children, since the children's level of maturity changes more quickly than adults. Most studies equalize the children's age, with their level of expertise. It is not obvious how children's ages influence results of a usability test.

Secondly, level of acquaintance is another important aspect in constructive interaction. Previous studies have indicated that children behave quite differently according to how well they know each other. In a study where adult test subjects were asked to bring a friend, co-worker, or family member to the usability test provided a positive experience [16] while other studies stress the importance of using non-acquainted test subjects [17]. Most studies involving children seems to prefer acquainted pairs of children; this is often achieved through involvement of children attending same school classes or kindergartens [10, 25, 28]. In the Eco-I project [30], the pairing goes beyond acquaintance, since a participating teacher had configured the pairs of children according to how well they worked together. Few studies indicate that the pairs of children were unacquainted, but this might have been the case in the StoryMat project [5] since the children attended different schools.

Thirdly, gender is potentially important when working with children; for example illustrated by girls and boys preferring different types of computer games [12]. Gender can also play a subjective role with children's preferences and attitudes towards technologies [4, 14]. But it is not apparent if and how gender influences other other issues of usability testing, such as effectiveness, efficiency, or number of identified usability problems. Several studies involve both genders in the design processes [3, 6, 19, 20, 30, 32, 33]. Some studies adapted imbalanced numbers of girls and boys [2, 25], while others deliberately chose an equal number of boys and girls [19]. Furthermore, some studies intentionally use same-sex pairs [10, 24].

Analyzing previous research on interaction design and children, we found several studies in which children participated as test subjects applying think-aloud [7, 8, 9, 28, 33], constructive interaction [24, 25, 30, 31], or both approaches [2, 6]. However, none of these studies present results related to how well the children adapted to think-aloud or constructive interaction. Summarized, we need a deeper understanding of involving children in the evaluation of software products to assess some of the opportunities and limitations related the different evaluation methods.

3 Experimental Method

The purpose of our experiment was to explore the impact of involving children in the evaluation of a software product. The idea was to place children in different settings

or conditions to see how this affects their performance. Thus, in this paper we do not measure the performance of the different setups in terms of usability problem identification (please refer to [1] for this aspect of our study).

Table 1: 60 children participated in our experiment in three different setups: constructive interaction as acquainted dyads or non-acquainted dyads and think-aloud as individual testers.

	Constructive Interaction		**Think-Aloud**
	Acquainted Dyads (N=24)	**Non-Acquainted Dyads (N=24)**	**Individual Testers (N=12)**
Girls	6x2	6x2	6
Boys	6x2	6x2	6
Total	12x2	12x2	12

We designed the experiment as a 3x2 matrix consisting of three types of sessions: individual testers using think-aloud, acquainted dyads (pairs) using constructive interaction, and non-acquainted dyads using constructive interaction. Furthermore, we configured the usability test sessions with same-sex dyads having sessions with girls and boys for each of the three setups. This is illustrated in table 1.

3.1 Participants

60 children (30 girls and 30 boys) at the age of 13 and 14 years old (M=13.35, SD=0.48) participated as test subjects in the experiment. The children were all 7th grade pupils from five different elementary schools in the greater Aalborg area. The children did not receive compensation for their involvement in the experiment.

The children were assigned as test subjects to one of the three test setups e.g. individual testers, acquainted dyads, or non-acquainted dyads. Each setup had twelve individual testers (six girls and six boys), twelve acquainted dyads (six pairs of girls and six pairs of boys), and twelve non-acquainted dyads (six pairs of girls and six pairs of boys). Assignment of the children to the three test setups was done randomly under two conditions 1) all acquainted dyads attended the same school class and 2) all non-acquainted dyads attended different schools. The acquainted pairs had known each other for at least five years except for one pair of girls and one pair of boys who had been acquainted for one year (M=6.25, SD=2.5). None of the non-acquainted dyads knew each other in advance.

3.2 System

The selected system for our experiment was an inno-100 mobile phone by innostream. This particular mobile phone was selected since it had not been released on

the European market at the time of our experiment. Thus, all children would have to learn to use the mobile phone.

The inno-100 integrates a range of standard mobile phone features, such as making and receiving phone calls and short text messages, and more advanced features, including speed dial functions and options for creating personalized ring tones. The inno-100 has two separate screens with a main 128x144 pixel 16 bit colour screen and 64x80 pixel sub screen on the cover. The navigation is primarily based on icons in the two upper menu levels. The lower levels are textual based including choice menus for setting values. Furthermore, the inno-100 integrates a number of games.

3.3 Procedure

Children from five schools in Aalborg, Denmark were introduced to the experiment by two of the participating researchers. The researchers explained the children's roles in the experiment and how their participation would contribute to our research. Participation in the experiment was voluntarily and interested children got an information sheet describing the experiment in detail and a consent form that had to be signed by a parent or a guardian. After receiving signed consent forms from a total of 60 children, we scheduled the usability evaluation sessions.

The sessions were held at the usability laboratory at Aalborg University. We adapted the guidelines for usability testing with children proposed by Hanna et al. [13]. Particularly, we focused on greeting the children, stressing the importance of the participation, and stressing that they were not the object of the test. The purpose of the evaluation was explained in detail to the children and they were shown the facilities of the usability lab. Test subjects intended for roles as non-acquainted dyads were kept separate before the test sessions. The children received questionnaires on which they had to provide answers to such as age, name, school, and mobile phone experience. The usability test sessions were conducted in a specialized usability laboratory. The laboratory integrated two rooms; an observation room in which the evaluations took place and a control room where one of the researchers would handle electronic equipment for recording the sessions. The two rooms were separated with a one-way mirror allowing people in the control room to see what was going on in the observation room. All sessions were recorded on video tapes for later analyses including perspectives of the children and of their interactions with the mobile phone.

The children were asked to solve twelve tasks one at a time addressing standard and advanced functionalities in the inno-100 mobile phone. This included making a phone call, sending a short text message, adjusting the volume of ring tones, and editing entries in the address book. We did not specify any time limits for the tasks, but required the participants to try to solve all tasks. All children were able to solve all specified tasks. On average, the children spent 26:45 minutes (*SD*=06:39) on the twelve tasks. The individual testers were asked to think-aloud while solving the tasks. We explained think-aloud to the individual testers in terms of the descriptions in [26, p. 195-198]. The acquainted and non-acquainted dyads were asked to solve the tasks by constructive interaction where they should collaborate with each other in order to

solve the tasks. We explained constructive interaction to the dyads in terms of the descriptions in [26, p. 198].

After the usability sessions, the children completed a subjective workload test (NASA-TLX) [22]. The children filled in the test individually even though the participated in pairs. This was done to evaluate the workload as experienced by the children in order to compare the different setups. We translated the test into the children's native language, Danish.

3.4 Data Analysis

After conducting all 36 sessions, the sessions were analyzed in a collaborative effort between two of the authors of this paper. The sessions were picked randomly for the analysis to avoid bias in the analysis. We analyzed the sessions according to how well the children collaborated (in constructive interaction sessions) and recorded their verbal interaction and comments. The six different aspects of our analysis were: 1) Level of verbalization, 2) quality of verbalization, 3) interaction between test subject(s) and test monitor, and 4) influence of test monitor on the solving of tasks. The two setups of constructive interaction were additionally analyzed according to: 5) Level of collaboration between the dyads and 6) quality of the collaboration between the dyads. We analyzed and marked each of the six aspects on a scale from 1 to 5 where 1 being the lowest score and 5 being the highest score. For example, for the level of verbalization, a session was marked 1 if the children made none or very few verbalizations during their interaction with the system, and a sessions was marked 5 if the children constantly or almost all time made verbalization during interaction.

The NASA-TLX tests were further analyzed. 55 tests were answered correctly by the children while 5 were incomplete answered. Data from our assessment of think-aloud and collaboration and the NASA-TLX tests were analyzed with one-way ANOVAs, followed by post hoc comparisons using Tukey tests.

4 Results

The 60 children in the 36 usability test sessions solved all 12 assigned tasks. Even though the constructive interaction sessions with acquainted dyads (M=29:54, SD=06:57) spent most time on the assignments in our experiment; the individual testers (M=25:34, SD=03:44), and the non-acquainted dyads (M=24:48, SD=07:53), we found no significant differences for the task completion times. The children performed and behaved differently in the three setups and the following sections present our assessment of their interaction and collaboration and the NASA-TLX test.

4.1 Assessment of Think-Aloud and Collaboration

As a part of our assessment of the three setups, we applied six different aspects of the verbalization and collaboration in usability tests. These six aspects are illustrated in

table 2. Not surprisingly, we found that the level of verbalization was considerably higher for the constructive interaction sessions compared to the think-aloud sessions. The acquainted dyads scored rather high (M=4.58, SD=0.90) especially compared the individual testers who scored rather low (M=2.17, SD=1.19). An analysis of variance shows significant differences between the three setups on level of verbalization $F_{(2,33)}$=13.421, p=0.001. A post-hoc test showed significant difference at the 0.1% level between the acquainted dyads and the individual testers and at the 5% level between the non-acquainted dyads and the individual testers. Furthermore, we found a tendency towards a higher level of verbalization for the acquainted dyads compared the non-acquainted dyads, but this difference is not significant (p=0.090).

Looking further at verbalization in the test sessions, we analyzed the quality of the verbalization primarily defined as the ability of the verbal comments to facilitate the identification and classification of usability problems. Considering the quality of the verbalization the differences between the setups are less apparent than for the level of verbalization. For the acquainted dyads (but also for some non-acquainted dyads), several verbal comments did not concern the actual test; a lot of the verbal comments did not facilitate the identification of usability problems. Summarized, the differences between the setups on quality of verbalization were not significant $F_{(2,33)}$=2.171, p=0.130.

Table 2: Assessment of verbalization and collaboration in the three setups. A plus indicates a significant difference to the setup marked with a minus according to an ANOVA test.

	Constructive Interaction		**Think-Aloud**
	Acquainted Dyads (N=12)	**Non-Acquainted Dyads (N=12)**	**Individual Testers (N=12)**
Level of verbalization	4.58 (0.90) +	3.58 (1.31) +	2.17 (1.19) -
Quality of verbalization	3.58 (1.00)	3.25 (1.48)	2.50 (1.38)
Interaction between test subject(s) and monitor	2.75 (0.87)	3.08 (0.79)	3.25 (0.87)
Influence of test monitor on the solving of tasks	2.17 (0.39)	1.67 (0.65)	1.83 (0.58)
Level of the collaboration between the dyads	4.75 (0.62)	3.83 (1.47)	N/A
Quality of the collaboration between the dyads	3.67 (1.56)	3.58 (1.56)	N/A

We further analyzed the influenced of and interaction with the test monitor. Constructive interaction provides potentially natural thinking-aloud as test subjects collaborate in pairs to solve tasks and therefore, one could expect less influence and interaction with a test monitor. We found that the test monitor has slightly more interaction with the think-aloud subjects compared the constructive interaction subjects, but the difference is not significant $F_{(2,33)}$=0.134, p=0.875. On the other hand, we identified a higher influence form the test monitor on the solving of tasks for the acquainted dy-

ads compared both non-acquainted dyads and individual testers, but again this difference is not significant $F_{(2,33)}$=0.282, p=0.756.

As constructive interaction have test subjects collaborate in pairs to solve tasks, we finally assessed the level and quality of collaboration. Most of the acquainted dyads collaborated during the entire sessions (*M*=4.75, *SD*=0.62) and we identified a tendency towards a higher collaboration between them than the non-acquainted dyads (*M*=3.83, *SD*=1.47), but this difference is not significant according to a Student's t-test $t_{(22)}$=1.993, p=0.059. Considering the quality of the collaboration, we found no difference between the two setups $t_{(22)}$=0.131, p=0.897.

4.2 Assessment of Subjective Workload

Table 3 summarizes mean values for the six factors of the NASA-TLX test as assessed by the 60 children in the three setups. As the table illustrates, minor differences could be observed between the different setups, however we found no significant differences between them. Even though not significant, we can however see that the individual testers found the effort factor more important than the dyads, but large variances were identified for the individual testers on this factor.

Table 3: Subjective workload (NASA-TLX test) for think-aloud and constructive interaction illustrating the mean values for the six factors as assessed by children.

	Constructive Interaction		Think-Aloud
	Acquainted Dyads (N=20)	Non-Acquainted Dyads (N=24)	Individual Testers (N=11)
Effort	38.5 (19.7)	41.9 (20.3)	52.7 (23.8)
Frustration	34.3 (25.4)	35.8 (22.4)	39.5 (23.4)
Mental	43.5 (16.2)	42.1 (19.3)	50.0 (12.2)
Performance	27.0 (21.7)	25.8 (17.7)	35.0 (24.5)
Physical	41.0 (25.8)	39.4 (25.9)	27.3 (13.8)
Temporal	38.5 (20.1)	27.5 (18.9)	37.7 (25.7)

On the other hand, more factors were assessed to almost the same mean values for the three setups e.g. frustration and mental demand. While the absolute values of the factors provided no significant differences between the three setups, we analyzed the inter-relative importance of the factors.

The assessment of the relative importance of the factors (table 4) showed significant difference between the three setups on the effort factor $F_{(2,52)}$=5.693, p=0.006. A post-hoc comparison showed significant difference at the 1% level between the acquainted dyads and non-acquainted dyads and at the 5% level between the acquainted dyads and the individual testers. Additionally, sitting with an acquainted influenced the importance of performance as acquainted dyads found this significantly more important than the individual testers and the non-acquainted dyads $F_{(2,52)}$=3.775,

p=0.029. A post-hoc test showed significant difference at the 5% level between the acquainted and non-acquainted dyads.

Table 4: Inter-relative assessment of workload factors for the three setups. A plus indicates a significant difference to the setup marked with a minus according to an ANOVA test.

	Constructive Interaction		Think-Aloud
	Acquainted Dyads (N=20)	**Non-Acquainted Dyads (N=24)**	**Individual Testers (N=11)**
Effort	2.30 (1.17) -	3.38 (1.10) +	3.45 (1.29) +
Frustration	1.75 (1.25)	2.54 (1.59)	2.64 (0.92)
Mental	2.90 (1.48)	3.38 (1.28)	3.73 (1.49)
Performance	3.15 (1.60) +	2.08 (1.18) -	2.09 (1.38)
Physical	2.35 (1.66)	2.08 (1.79)	1.36 (1.75)
Temporal	2.55 (1.50)	1.54 (1.47)	1.73 (1.27)

We found that the acquainted dyads assessed frustration as the least important factor while both individual testers and non-acquainted dyads rated it as the third most important factor, but this difference was not significant $F_{(2,52)}=2.337$, $p=0.107$. For the remaining three factors, we found only minor differences between the three setups and no significant differences, mental demand $F_{(2,52)}=1.357$, $p=0.266$, physical demand $F_{(2,52)}=1.160$, $p=0.322$, while we identified a tendency for temporal issues $F_{(2,52)}=2.800$, $p=0.070$.

Table 5: Calculated workload for the three setups. A plus indicates a significant difference to the setup marked with a minus according to an ANOVA test.

	Constructive Interaction		Think-Aloud
	Acquainted Dyads (N=20)	**Non-Acquainted Dyads (N=24)**	**Individual Testers (N=11)**
Effort	99.8 (80.0) -	148.7 (90.7)	190.9 (126.3) +
Frustration	61.0 (63.1)	108.8 (97.5)	116.8 (91.5)
Mental	120.8 (65.1)	132.5 (69.5)	186.8 (90.7)
Performance	83.8 (88.4)	51.7 (56.0)	65.0 (50.0)
Physical	118.8 (113.8)	80.0 (104.4)	40.5 (61. 6)
Temporal	90.0 (60.7) +	42.1 (55.6) -	58.2 (59.3)

Combining the two measures, we calculated the overall score for the workload for the participating children. As discovered above, we found that the individual testers had to put much more effort into the testing situation and an ANOVA test showed a significant difference between the three setups $F_{(2,52)}=3.464$, $p=0.039$. A post-hoc comparison showed significant difference at the 5% level between the individual testers

and the acquainted dyads. On the other hand, the acquainted dyads in total assessed temporal demand rather high compared to the two other setups and we found a significant difference between the three setups $F_{(2,52)}=3.737$, $p=0.030$. A post-hoc test showed significant difference at the 5% level between the acquainted dyads and the non-acquainted dyads.

We identified a tendency for mental demand as the individual testers in general assessed this factor higher than both constructive interaction setups, however the difference was not significant for our test $F_{(2,52)}=3.114$, $p=0.057$. Again and as above, we found that the level of frustration is much lower for the acquainted dyads compared the two other setups, however the difference is not significant $F_{(2,52)}=2.247$, $p=0.116$. Furthermore, we found no significant differences for the other calculated values; physical demand $F_{(2,52)}=2.198$, $p=0.121$ and performance $F_{(2,52)}=1.190$, $p=0.312$.

5 Discussion

This section provides qualitative results from the study. We have identified a number interesting lessons related usability testing with children.

Lesson 1: *Constructive interaction did not necessarily facilitate natural think-aloud as the dyads tended to talk-aloud and not think-aloud*. Constructive interaction in usability testing with children potentially provides natural thinking-aloud as the children collaborate in pairs to solve tasks. Our study illustrated that children in pairs using constructive interaction had a much higher level of verbalization, but often they were more talking-aloud than actually thinking-aloud. We further experienced that the individual testers applying think-aloud tended to be quieter during the sessions compared to the dyads; they expressed themselves noticeably fewer times than the dyads. When asked about their choices, more of them would mostly answer our questions in very few words without giving further insight into their behaviour and choices. On the other hand, the non-acquainted dyads had less interaction with each other compared to the acquainted dyads; they mainly kept focus on the task they were solving. The interaction of the acquainted dyads was partially related to the task, but we identified some interaction as noise as this was irrelevant to the solving of the task, for example some would have long discussions on what to name the melody they had just composed. These observations resemble the discussion by Ericsson and Simon of think-aloud and talk-aloud [11]. It was very difficult to get the children to explain their interaction and motivation even though they had been carefully instructed before the session. Thus, this can be seen as a contradiction to benefits of constructive interaction as stated by Nielsen [26] as we found only minor differences between the think-aloud sessions and constructive interaction sessions.

Lesson 2: *Dyad configuration in constructive interaction influenced the children's behaviour and assessment of the testing situation according to their acquaintance*. Our study indicated that there were a significant difference between how the acquainted and the non-acquainted dyads experienced the assessment of effort and performance. Our results showed that the acquainted dyads were significantly more satisfied with their own performance and they did not feel it demanded a lot of effort from them. It was just the opposite for the non-acquainted dyads. Even though the

acquainted dyads sometimes would try to pull the phone out of the hands of their co-solver, they rated performance of minor importance compared to the non-acquainted dyads. From our study, we also found that the non-acquainted dyads acted rather polite against each other and in general they were more polite to each other than the acquainted dyads. Consequently, they collaborated quite differently compared to the acquainted dyads and they did not argue explicitly for the control of the tested phone. This is also indicated in our results as we found a tendency, however not significant, towards better collaboration between the non-acquainted dyads. Further, the non-acquainted dyads separated the roles between them during the test. Even in the cases were they did not collaborate very well, they would some times read the task aloud, or they would take turns by shifting in between tasks. The acquainted dyads' interaction were influenced by the fact that the children knew each other in advance, they referred to each others by nick-names, remarked their co-solvers intelligence etc. They would also physically try to grab the phone and thereby preventing their co-solver from helping to solve the task. The acquainted dyads would easily get distracted from the task they were solving, they would discover something interesting in the menu, and would spend time discovering such aspects. Some of the non-acquainted dyads did not collaborate very well while solving the task; we found no significant differences between the girls and the boys in this issue. The children took turns in operating the system and the child who was not in control of the interaction had sometimes difficulties in seeing what was going on the screen of the phone.

Lesson 3: *Gender issues might play important roles in the configuration of dyads in constructive interaction.* Our study utilized pairs of same sex dyads as adapted in several studies with children [10, 24]. Even though we haven't summarized the results gender wise, our study showed a tendency towards that the boys collaborated better than the girls. Especially the acquainted dyads of boys collaborated rather well and had a fruitful and successful collaboration whereas the acquainted dyads of girls experienced several situations where their collaboration was rather poor. Thus, while it seems to be of less importance if the boys tested in acquainted or non-acquainted dyads, the girls should test in non-acquainted dyads. For some of the specified tasks, we observed that the acquainted dyads of girls would more easily get distracted from the task they were solving, they would discover something interesting in the menu, and would spend time discovering what it was, for example acquainted dyads quite often used several minutes to compose a melody, for example "Itsy Bitsy Spider".

6 Conclusion

In this paper, we investigate and address the above stated questions by looking at how children perform and behave in constructive interaction during usability testing. Our particular focus is on how the children behave and perceive a testing situation when involved a traditional think-aloud test compared to constructive interaction tests. Thus, we did not treat the performance of the different setups in terms of usability problem identification (please refer to [1] for this aspect of our study).

Our results show that the pairing of children had impact on how the children verbalized and collaborated in pairs during the testing sessions. First, we found that constructive interaction did not necessarily facilitate natural think-aloud as the dyads tended to talk-aloud and not think-aloud. Our children in pairs had a high level of verbalization, but often they were more talking-aloud than actually thinking-aloud. This issue resembles some of the discussions by Ericsson and Simon of think-aloud and talk-aloud [11]. Secondly, dyad configuration in constructive interaction influenced the children's behaviour and assessment of the testing situation according to their acquaintance. The acquainted dyads were significantly more satisfied with their own performance and they did not feel it demanded a lot of effort from them. It was just the opposite for the non-acquainted dyads. Thirdly, gender issues might play important roles in the configuration of dyads in constructive interaction. Our study showed a tendency towards that the boys collaborated better than the girls. Especially the acquainted dyads of boys collaborated rather well and had a fruitful and successful collaboration whereas the acquainted dyads of girls experienced several situations where their collaboration was rather poor. Thus, while it seems to be of less importance if the boys tested in acquainted or non-acquainted dyads, the girls should test in non-acquainted dyads.

Our study suffers from a number of limitations which could form further research with children. First, our results of our experiment cannot simply be generalized for all ages of children. Thus, replicating the experiment with younger children may show a different kind of relationship between think-aloud and constructive interaction. Secondly, we recorded that the non-acquainted dyads continuously took turns with the mobile phone making it difficult for the other child to see what was going on at the interface. This could probably be different for desktop-based applications.

Acknowledgements

The work behind this paper received financial support from the Danish Research Agency (grant no. 2106-04-0022). We would especially like to thank all the participating children and their parents. Furthermore, we want to thank several anonymous reviewers for comments on drafts of this paper.

References

1. Als, B. S., Jensen, J. J., and Skov, M. B. (2005) Comparison of Think-Aloud and Constructive Interaction in Usability Testing with Children. In *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, ACM Press
2. Benford, S., Bederson, B. B., Åkesson, K-P, Bayon, V., Druin, A., Hansson, P., Hourcade, J. P., Ingram, R., Neale, H., O'Malley, C., Simsarian, K. T., Stanton, D., Sundblad, Y., and Taxén, G. (2000) Designing storytelling technologies to encouraging collaboration between young children. In *Proceedings of the Human Factors and Computing Systems CHI'00*, ACM Press, pp. 556 - 563

3. Bers, M. U., Gonzalez-Heydrich, J., and DeMaso, D. R. (2001) Identity Construction Environments: Supporting a Virtual Therapeutic Community of Pediatric Patients Undergoing Dialysis. In *Proceedings of the Human Factors and Computing Systems CHI'01*, ACM Press, pp. 380 - 387
4. Cassell, J. (2002) *Genderizing*. The Handbook of Human-Computer Interaction
5. Cassell, J. and Ryokai, K. (2001) Making Space for Voice: Technologies for Supporting Children's Fantasy and Storytelling. *Personal and Ubiquitous Computing*, Springer-Verlag, vol. 5(3), pp. 169 - 190
6. Danesh, A., Inkpen, K. M., Lau, F., Shu, K., Booth, K. S. (2001) Geney: Designing a collaborative activity for the Palm handheld computer. In Proceedings of the Human Factors and Computing Systems CHI'01, ACM Press, pp. 388 - 395
7. Druin, A. and Solomon, C. (1996) Designing Multimedia Environments for Children. Wiley & Sons, New York
8. Druin, A. (1999) The Role of Children in the Design of New Technology. HCIL Technical Report No. 99-23, University of Maryland, USA
9. Druin, A. (1999) The Design of Children's Technology. Morgan Kaufmann Publishers, Inc., San Francisco, CA
10. Ellis, J. B. and Bruckman, A. S. (2001) Designing Palaver Tree Online: Supporting Social Roles in a Community of Oral History. In Proceedings of the Human Factors and Computing Systems CHI'01, ACM Press, pp. 474 - 481
11. Ericsson, K.A. and Simon, H.A. (1990) *Protocol Analysis. Verbal reports as data*, Cambridge Massachusetts
12. Gorriz, C. M. and Medina, C. (2000) Engaging Girls with Computers through Software Games. Communications of the ACM, vol. 43, No. 1, pp. 42 – 49
13. Hanna, L., Risden, K., and Alexander, K. J. (1997) Guidelines for Usability Testing with Children. In interactions, September + October, pp. 9 – 14
14. Inkpen, K. (1997) Three Important Research Agendas for Educational Multimedia: Learning, Children, and Gender. In Proceedings of Educational MultiMedia '97
15. Jensen, J. J. and Skov, M. B. (2005) A Review of Research Methods in Children's Technology Design. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05), ACM Press
16. Kahler, H. (2000) Constructive Interaction and Collaborative Work. interactions, May + June, pp. 27 - 34
17. Karat, C.-M., Campbell, R., and Fiegel, T. (1992) Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. In Proceedings of the Human Factors and Computing Systems CHI'92, ACM Press, pp. 397-404
18. van Kesteren, I. E. H., Bekker, M. M., Vermeeren, A. P. O. S., and Lloyd, P. A. (2003) Assessing usability evaluation methods on their effectiveness to elicit verbal comments from children subjects. In Proceeding of the 2003 conference on Interaction design and children (IDC'03), ACM Press, pp. 41 - 49
19. Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., and Bhogal, R. S. (1997) The Persona Effect: Affective Impact of Animated Pedagogical Agents. In Proceedings of the Human Factors and Computing Systems CHI'97, ACM Press, pp. 359 - 366
20. Lumbreras, M. and Sánchez, J. (1999) Interactive 3D Sound Hyperstories for Blind Children. In Proceedings of the Human Factors and Computing Systems CHI'99, ACM Press, pp. 318 - 325
21. Markopoulos, P. and Bekker, M. (2003) On the Assessment pf Usability Testing Methods for Children. Interacting with Computers, Elsevier, Vol. 15, pp. 227 – 243

22. Miller R. C. and Hart, S. G. (1984) Assessing the Subjective Workload of Directional Orientation Tasks. In Proceedings of 20th Annual Conference on Manual Control, NASA Conference Publication, pp. 85 – 95
23. Miyake, N. (1986) Constructive Interaction and the Iterative Process of Understanding. Cognitive Science, vol. 10(2), pp. 151 - 177
24. Moher, T., Johnson, A., Ohlsson, S., and Gillingham, M. (1999) Bridging Strategies for VR-based Learning. In Proceedings of the Human Factors and Computing Systems CHI'99, ACM, pp. 536 - 543
25. Montemayor, J., Druin, A., Farber, A., Simms, S., Churaman, W., and D'Amour, A. (2002) Physical Programming: Designing Tools for Children to Create Physical Interactive Environments. In Proceedings of the Human Factors and Computing Systems CHI'02, ACM Press, pp. 299 - 306
26. Nielsen, J. (1993) Usability Engineering. Academic Press
27. O'Malley, C. E., Draper, S. W., and Riley, M. S. (1984) Constructive Interaction: A Method for Studying Human-Computer-Human Interaction. In Proceedings of IFIP Interact '84, pp. 269 – 274
28. Rader, C., Brand, C., and Lewis, C. (1997) Degrees of Comprehension: Children's Understanding of a Visual Programming Environment. In Proceedings of the Human Factors and Computing Systems CHI'97, ACM Press, pp. 351 - 358
29. Resnick, M., Martin, F., Berg, R., Borovoy, R., Colella, V., Kramer, K., and Silverman, B. (1998) Digital Manipulatives: New Toys to Think With. In Proceedings of the Human Factors and Computing Systems CHI'98, ACM, pp. 281 – 287
30. Scaife, M., Rogers, Y., Aldrich, F., and Davies, M. (1997) Designing for or Designing with? Informant Design for Interactive Learning Environments. In Proceedings of the Human Factors and Computing Systems CHI'97, ACM Press, pp. 343 - 350
31. Skov, M. B., Andersen, B. L., Duhn, K., Garnæs, K. N., Grünberger, O., Kold, U., Mortensen, A. B., and Sørensen, J. A. L. (2004) Designing a Drawing Tool for Children: Supporting Social Interaction and Communication. In Proceedings of the Australian Computer-Human Interaction Conference 2004 (OzCHI'04)
32. Stewart, J., Bederson, B. B., and Druin, A. (1999) Single Display Groupware: A Model for Co-Present Collaboration. In Proceedings of the Human Factors and Computing Systems CHI'99, ACM, pp. 286 - 293
33. Strommen, E. (1998) When the Interface is a Talking Dinosaur: Learning across Media with ActiMates Barney. In Proceedings of the Human Factors and Computing Systems CHI'98, ACM Press, pp. 288 - 295

A Case Study of Three Software Projects: Can Software Developers Anticipate the Usability Problems in their Software?

*Rune Thaarup Høegh† and Janne Jul Jensen†

†Aalborg University, Department of Computer Science,
Frederik Bajers vej 7E, DK-9210 Aalborg East, Denmark

*Corresponding author. Email: runethh@cs.aau.dk

Abstract

The purpose of usability evaluations is typically to discover which areas of a system that perform satisfactory to the end-user and which areas that need redesigning or improving. However, such evaluations can be costly both in time and funds and when developers say that many of the results from the usability evaluations are issues already known to them, then why bother? This paper discusses the result of three case studies in which the participants of a development process were asked to describe the usability problems of the system they had helped develop. These descriptions were then compared to the results of a usability evaluation involving end-users to uncover if software developers can describe which usability problems exist in their software. To some extent they can. However, they do not always agree on the problems, and the severity ratings were often different from the ones based on the experiences from the users. Furthermore, the developers' description of the problems was typically more abstract and less detailed than the descriptions from the usability evaluation. The tendency was that the most critical problems and the problems most often experienced were listed by the participants and thus the amount of problems known by the developers was a lot less than the amount of problems discovered by the usability evaluation.

Keywords: Usability evaluation, Human Computer Interaction.

AMS Subject Classification: 68N99;

1 Introduction

Usability evaluations are applied to assess the quality of a user interaction design in a software system and establish a basis for improving it (Rubin, 1994). This is accomplished by identifying specific parts of a system that do not properly support the users in carrying out their work. Thus usability evaluations and the related activities can help developers make better decisions, and thereby allow them to do their jobs more effectively (Radle and Young, 2001). The result of usability evaluations is often accentuated as a distinctive input for developers to improve the usability of a software system. On the other hand developers say that many of the results from the usability evaluations are issues already known to them. This study examines the amount and nature of usability problems developers are aware of prior to a usability evaluation, in order to emphasize the type of usability problems that still needs pointing out. The study mainly involves project participants that have a direct relation to the graphical user interface, either because they are developing them, or because they are supporting or teaching about them. Back-end developers are not included in this study for practical purposes, although it can be argued that they also have influence on the graphical user interface.

2 Related Work

Card, Moran and Newell (1983) suggested the concept of mental models of interactive computer systems. The idea of mental models came out of cognitive science, and was also supported by Norman (1983).
The mental models refer to a number of models, including the model of the system, the user interface designer's mental model of the system, and the user's mental model of the system. This idea allowed for future HCI researchers a framework for understanding the ease of use for a particular design.

One could argue that every usability evaluation is a study in comparing the developer's models of a design to the user's mental model; it is however not often in the HCI literature that the developers' knowledge of usability problems in their own software is described. Studies of the User Centered Design (UCD) approach (eg. Greenbaum and Kyng, 1991) report from studies where designers have had their assumptions about a designs usability tested (e.g. Alexander 2003, Van House et al 1996). The experiences from projects developed by UCD show that developers often get surprised by seeing that users can not use the software the way it has been designed. The general conclusion from UCD studies is that software where the developers test their assumptions by using UCD techniques has a higher chance of resulting in usable software (Mao et al, 2005). These experiences show that it is often the case that developers do not have an adequate feeling for which part of a design that may be user friendly.

In contrast to this are the results of (Høegh et al, 2006). They describe three different relationships between usability evaluators and developers. The first relationship is the situation where the evaluator and the developer are integrated in the same development team, and the evaluator and the developer may even be the same person. The second relationship is when the evaluators form a separate organizational unit within the development organization and the third relationship is where evaluators are employed by a different organization. The dominant form of relationship for development projects reported in the HCI literature appears to be the first type of relationship. In a recent literature review of papers presenting studies that included usability evaluations reported that 81 percent of the development projects where organized with developers and evaluators being integrated in the same development team or even the same person (Høegh, 2006). The above described picture of software development do not inform about software projects that did not include usability evaluations. The authors' assumption is however that a great deal of software is being developed without ever being subject to a user based usability evaluation, hence the developers' knowledge of the usability problems becomes the predominant source of information to correct usability problems.

Thus, in spite of research suggesting that developers have an inadequate sense of the usability of their own software, still in many cases it is those very same developers who test the usability of the software. Given this situation it underlines the importance of finding out exactly which usability problems the developers are able to identify and which are only identified by a user evaluation.

3 Case study

This section describes in detail the case study involving all three software projects. It covers participants, procedure and the software. The purpose of the study was to expose which usability problems the participants knew in advance and which usability problems were more hidden to the participants and needed to be exposed with an evaluation instead.

3.1 Participants

Software developers from three software projects were included in this case study. All three software projects were large commercial projects. Three software developers participated from project A and three from project B. From project C two software developers, a supporter, an educator and a manager was involved. All software developers had a masters degree or similar in Computer Science, and had worked on the software projects through most of the development. Project C furthermore involved a member from the support staff, one of the educators who teach the clients to use the software and the section leader.

3.2 The Three Software Projects

Each of the three software projects had been running for more than two years, and the software in the projects had working and functional graphical user interfaces. The company behind project A and B develops software to the telecommunication industry, and the company behind project C develops software to the healthcare industry. Project C is already in use out in the field and has been for several years. It is under continuous development.

3.3 Procedure

The software was subjected to a usability evaluation with users. For each software project, users with relation to the software's domain were involved. The users were asked to solve a number of tasks related to the intended future use of the software, and they were asked to think aloud during the task solving. The users' actions were recorded on video, and the video was afterwards analyzed in order to identify usability problems experienced by users. The analysis was done by the authors, who each have several years of experience in the HCI field, and have conducted several usability evaluations before.

Separate from the usability evaluation and analysis, a workshop was held for each of the projects. For each workshop, the people involved in the project were invited. The overall purpose of the workshop was to discuss the participants' impressions of what usability problems the software contained. However, the setup for project C differed slightly from the setup of projects A and B. The focus of setup A and B was to reveal the total number of usability problems the developers knew, whereas in project C, the setup was aimed at finding what the developers perceived to be the most severe usability problems.

In both project A and B the participants were informed which tasks that were used for the usability evaluation with the users, and a description of what constituted a usability problem was presented. This was done to focus the attention of the developers to the area of the usability evaluations that were evaluated, and to ensure that the authors and the developers had a common understanding of what usability was. For each software project the participants were then each asked to individually write down all known usability problems in the software they have been part of developing. They were asked to assess the severity of each problem as well as write in detail how and when the problem occurred and why they thought it was a problem. The individual lists were then merged into a common list. The usability problems identified in the usability evaluation and the participants' lists of usability problems were then compared in relation to the amounts of usability problems, the nature of the usability problems, and the severity of the usability problems.

In project C, the participants were not informed about the tasks from the usability evaluation. The developers were instead asked to write down the most severe usability problems they expected to be in the system as a whole. Furthermore, project C involved not only developers, but a range of professions that had all been a part

of the development process. Unlike project A and B, the participants were not asked to write down all known problems but rather to name the top three problems they knew of (if any). This top three was based on their perceived seriousness of the problem. They too were asked to write in detail how and when the problem occurred and why they thought it was a problem. These top threes were then compared to the top problems of the list of usability problems from the evaluation.

4 Results

The results from the workshops and the usability evaluations are shown in table 1. In the three projects, the usability evaluation with users revealed 80, 70 and 105 usability problems respectively, and the developers of project A and B had combined named 14 and 22 usability problems. In project C the participants were asked for a top three, which lead to a merged list of 12 usability problems. This list would probably have been longer if the participants had been asked to list all known problems, and therefore should not be compared to the merged list from project A and B. It is interesting that the merged list from project A yields only two additional problems compared to developer Cs list. This implies that the developers were quite in agreement as to what usability problems existed. Opposite is project C where only one usability problem on the top threes had some agreement. This problem was named by three of the participants as being severe or critical and was also the most critical problem in the usability evaluation. However, apart from that one usability problem, the rest were all different usability problems, so there was almost no agreement between the participants as to what were the most severe usability problems. This is interesting since it suggests that the developers have a differing view on what is a usability problem compared to the support and educator who are interacting with the end users. Furthermore, the usability problems listed by the supporter and the educator were much more in agreement with the highest ranking usability problems of the usability evaluation. Finally project B lies between project A and C concerning agreement.

Table 1- The described and found usability problems for each project

Project A		Project B		Project C	
Source	Usability problems	Source	Usability problems	Source	Usability problems
Developer A	8	Developer D	8	Developer G	3
Developer B	10	Developer E	15	Developer H	1
Developer C	12	Developer F	4	Supporter	3
				Educator	3
				Section leader	4
Merged list	14	Merged list	22	Merged list	12
Usability evaluation	80	Usability evaluation	70	Usability evaluation	105

The results from the usability evaluations and their relation to the developers expected usability problems are displayed in table 2, 3 and 4. For each table, the first row represents the usability problems described by the developers and the second row presents the segment of the usability problems identified during the video based analysis of the usability evaluation that corresponds to a usability problem described by a developer. The usability problems found through the usability lists were more specific than those described by the developers. Hence the descriptions of usability problems in the first row list sometimes cover two or more specific usability problems seen in the usability evaluation. A ❶ denotes a critical problem, a ❷ denotes a serious problem and a ❸ denotes a cosmetic problem (Molich, 2000). In project C, the participants did not always rate the problem and this is denoted by ❼. Note that the second row does not show all usability problems identified in the usability evaluation, as not all problems corresponded to a usability problem

described by the developers. Finally, keep in mind that project C only listed the top problems and not all known problems.

Table 2- Relation between project A's list of usability problems from the usability evaluation and the developers list.

Developers list	❶	❶	❶	❶	❶	❷	❷	❷	❷	❸	❸	❸	❸	❸
Usability evaluation		❷❷❷❷ ❸		❶❶❷❷	❷	❶❷ ❸❸	❷			❷❷ ❸	❶	❷❷❷❷ ❸❸❸❸	❸	❷❷

In project A there were four usability problems listed by the developers that were not experienced by any users during the usability evaluation.

Table 3- Relation between project B's list of usability problems from the usability evaluation and the developers list.

Developers list	❶	❶	❶	❶	❶	❷	❷	❷	❷	❷	❸	❸	❸	❸	❸	❸	❸	❸	❸	❸
Usability evaluation	❷❷ ❸	❶❷ ❷❸	❶❶ ❸	❶❶ ❷	❷			❷				❷❷		❷	❶❷ ❸	❸	❷❸ ❸❸				

In project B 11 out of the 22 problems on the merged list were not experienced by any users. One of the problems was impossible to verify with only one simultaneous user of the system, as the problems was related to distribution of the workflow between several users. The remaining usability problems could have been experienced by users, but either the users did not experienced them, or they did not experience them as usability problems.

Table 4 – Relation between project C's list of usability problems from the usability evaluation and the developers top threes.

Developers list	❶	❸	❶	❶	⓿	⓿	⓿	⓿	⓿	⓿	❶	⓿
Usability evaluation	❶❶❶❶										❶❶❶❶❸❸ ❸❸	❶❶❷❷❷❷❸❸

In project C, only three of the 12 problems listed in the top threes were problems revealed by the usability evaluation. Three of the problems not experienced by the users in the usability evaluation were problems that could not be verified due to it e.g. requiring several users working simultaneously during the test, or server access not granted for the experiment. The remaining six usability problems could have been experienced by users, but was not.

What is evident in all three cases is that the participants typically are aware of several of the most critical problems in the software. In project C for instance, eight of the ten most serious problems exposed by the user evaluation were all known and named problems to several of the participants. However, they were formulated in a very general way and lacking the detail and concreteness of the descriptions from the usability evaluation, thus is registered as only three problems by the participants although they cover many more in the usability evaluation. Hence some of the usability problems written by the participants would cover two or more of the usability problems identified by the usability evaluation with users. And as could be suspected beforehand, especially the supporter and educator had a good feeling which problems were most critical to the users. Furthermore the participants tend to downplay the seriousness of a given problem. As an example in project C one developer describes a problem and then comments: "The users often don't have much experience with technology, but once the procedure and the 'rules' are explained to them, they should be able to overcome those types of problems". It is clear that this developer places the problem more as a learning or experience problem than an actual usability problem. Similarly the same developer comments: "This should be easily learned. I keep forgetting it myself though." Finally, it was also observed that the various participants named different usability problems; hence they did not have a shared meaning of what the usability problems in their

software were. This was not only related to project C, where the participants had different jobs in the development process. Also between the software developers in project B there was disagreement as to what problems existed. Only in project A, there seemed to be some consensus. In all three projects there was a tendency to rethink the seriousness of a problem when presented with actual video of the problem occurring to a user compared to just having the problem presented in writing in a report. There was a consensus that it became more convincing and clear that way and was harder to dismiss as being the users own fault or similar.

5 Discussion

In all the projects, the same tendency could be seen; the participants had knowledge of some of the usability problems prior to the evaluations, but they were however mostly only able to describe about a third of the usability problems according to project A and B. However, according to project C, the problems mentioned do comply with many of the most serious problems from the usability evaluation. The usability problems described by the developers where furthermore in more general terms than those identified during the usability evaluation. A usability problem on one of the merged lists was described as “Feedback on errors is not good enough”, whereas a similar usability problem was described as “The user can not read in the displayed error message (specific message code) why the system broke down”. The second description is a lot more specific than the first, as it refers to a specific situation and exemplifies what the problem is.

For all projects there were quite big differences between the usability severity rating given by the developers and the usability severity ratings given by the video based analysis. The comparison of severity rating were complicated by the difference in abstraction level, but it can never the less be observed that the usability evaluation of the software did provide the developers with a more accurate idea of what the severity of the usability problems is.

We can conclude that the usability evaluation added more specific knowledge about the state of the software projects, both in terms of the type of usability problems, the amount of usability problems, and the severity of the usability problems. The participants also listed usability problems that had not been experienced in the evaluation with users. Finally all three cases show a wide variety in the problems listed between the participants, which indicates that they have either a differing view on what constitutes a usability problem or simply a differing view on what actual usability problems the software contains.

In practice this means that regardless of the claims from the developers themselves, they do not know the problems of their own software, and thus if a more structured overview is desired, a usability evaluation is in order. However, the developers seem to be able to supply the most critical problems of the software, thus their input is definitely better than nothing. Finally, the participants listed problems not revealed by the usability evaluation and it would be interesting to look further into these, to verify if they indeed are usability problems.

6 Limitations

The study holds a few limitations that are worth taking into consideration when evaluating the results. The study is in essence comparing two types of usability evaluation methods (UEM); an expert review (or free recall) and a user based laboratory test based on tasks. The differences in these two UEMs of course limits the standard of reference, but the study were designed to minimize this, as a common understanding of usability was sought. A closer look at the comparison of the usability problem listed by the two approaches shows that only few problems described by the developers could not have been found in the usability test.

Acknowledgement

The research behind this paper was partly financed by the Danish Research Councils (grant number 2106-04-0022, the USE-project). Also thank you to the participating companies and the home healthcare workers of Aars kommune in Denmark, who agreed to participate in this experiment, and to the elderly citizens, who so willingly opened their homes to us.

References

ALEXANDER, D. 2003, Redesign of the Monash University Web Site: A Case study in User-Centered Design Methods. Proceedings of AusWeb 2003.

CARD, S.K., MORAN, T. and NEWELL, A. 1983. *The* Psychology of Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

GREENBAUM, J., KYNG, M. 1992, Design at Work: Cooperative Design of Computer Systems. Lawrence Erlbaum Associates.

HØEGH, R. T. The Focus of Current HCI Research in Usability Evaluation and Feedback. Proceedings of APCHI 2006.

HØEGH, R. T., NIELSEN, C, M., OVERGAARD, M., PEDERSEN, M. P. AND STAGE, J. 2006, A Qualitative Study of Feedback From Usability Evaluation to Interaction Design: Are Usability Reports Any Good? Special Issue of International Journal of Human Computer Interaction, Lawrence Erlbaum Associates.

MAO, J. VREDENBURG, K. SMITH, P. W., CAREY, T. 2005, The State of User-Centered Design in Practice. Communications of the ACM, March 2005, volume 48, 3. ACM press.

MOLICH, R. 2000. Brugervenlige EDB systemer. (Eng: User-Friendly Computer Systems). Teknisk Forlag.

NORMAN, D.A. 1983. Some observations on mental models. In Mental Models, edited by D.Gentner and A. Stevens. Lawrence Erlbaum Associates: Hillsdale, NJ.

RADLE, K. and YOUNG, S. 2001, Partnering Usability with Development: How Three Organizations Succeeded. IEEE Software, January/February 2001

RUBIN, J. 1994. Handbook of usability testing: How to plan, design, and conduct effective tests, New York, NY: John Wiley & Sons.

VAN HOUSE, N. A., BUTLER, M. H. OGLE and V. SCHIFF, L. 1996, User-Centered Iterative Design for Digital Libraries. D-Lib Magazine.

Composing children dyads in constructive interaction: a comparison of usability testing methods for problem identification

B. S. Als, J. J. Jensen, M. B. Skov

Department of Computer Science, Aalborg University
Selma Lagerlöfs Vej 300, DK-9220 Aalborg East, Denmark
{als, jjj, dubois}@cs.aau.dk

ABSTRACT. Constructive interaction provides natural thinking-aloud as test subjects collaborate to solve tasks. Since children may face difficulties in following instructions for a standard think-aloud test, constructive interaction has been suggested as evaluation method when usability testing with children. However, the relationship between think-aloud and constructive interaction is still poorly understood. We present an experiment that compares think-aloud and constructive interaction in usability testing. The experiment involves 60 children with three setups where children apply think-aloud, and constructive interaction in acquainted and non-acquainted pairs. Our results show that the pairing of children had significant impact on the identification of usability problems as acquainted dyads identified more problems than non-acquainted dyads. Furthermore, we found significant differences between constructive interaction and think-aloud for problem identification. Finally, the acquainted pairs reported less frustration during the test, despite the identification of more problems.

1. Introduction

The design and evaluation of children's technologies have received increased attention during the last several years (Druin, 1999b). Children should be considered individuals with strong opinions, needs, likes, and dislikes, and they should be treated as such (Druin and Solomon, 1996). When evaluating technologies with children, evaluators are typically faced with unique challenges as children enter usability evaluations with special preconditions (Hanna et al., 1997). Thus, we need to understand how to create successful environments for children that facilitates usability problem identification.

Comparative studies of usability evaluation methods (UEMs) have focused on the vast number of UEMs, and their opportunities and limitations in evaluating software products (Gray and Salzman, 1998; Jeffries et al., 1991; Karat et al., 1992). The think-aloud protocol has been credited for its effectiveness in identification of usability problems (Karat et al., 1992; Molich, 1997; Rubin, 1993). In usability testing with children, Nielsen (1993) suggests that evaluators should use a variation of think-aloud called constructive interaction (Miyake, 1986; O'Malley et al., 1984) also referred to as co-discovery learning (Kahler, 2000). The rationale behind this recommendation is that it may be difficult to get children to follow the instructions for a standard thinking-aloud test (Nielsen, 1993). Constructive interaction involves two test subjects collaborating in trying to solve tasks while using a computer system (O'Malley et al., 1984). However, we still lack empirical evidence of the merits of constructive interaction in usability evaluations with children.

In this paper, we investigate problem identification from think-aloud and constructive interaction in an experiment by comparing usability tests where children employ think-aloud and constructive interaction respectively while interacting with a mobile phone. We are especially concerned with two issues. First, we wish to compare think-aloud and constructive interaction on the number and types of usability problems identified. Secondly, we wish to explore the impact of pair composition on constructive interaction, i.e. the social relationship between the children (again against problem identification).

2. Related Work

Constructive interaction facilitates natural thinking-aloud as subjects interact and collaborate to solve tasks while interacting with the system (O'Malley et al, 1984). On the other hand, subjects in constructive interaction may aim for different strategies for learning and using computers. Furthermore, since constructive interaction requires twice as many test subjects as think-aloud, in order to conduct the same number of usability sessions, it is typically more expensive (Nielsen, 1993).

Composing pairs or dyads in constructive interaction introduces a number of issues to be considered for usability evaluators (Miyake, 1986). One key issue relates configuring pairs in constructive interaction on their level of expertise. The level of expertise is important, as argued by O'Malley et al. (1984), since the test subjects' knowledge of specific work tasks is quite often corresponding to their level of expertise. Nielsen (1993) recommends that the test subjects have the same level of experience, whereas having one of the test subjects able to guide the interaction, is an argument used by Kahler (2000) when stating advantages by pairing test subjects with different levels of experience. Usually children do not posses expertise of work that might influence the outcome of the usability test, which makes the issue of expertise more subtle when working with children. Most studies involving children do not explicitly consider the level of expertise (Lester et al., 1997; Montemayor et al., 2002), one of the exceptions is a study where the participating children are profiled according to their scripting level (Rader et al., 1997). Where age does not seem to matter when testing with adults, it has a more eloquent impact when conducting tests with children, since the children's level of maturity changes more quickly than adults. Most studies equalize the children's age, with their level of expertise. However, it is not obvious how children's ages influence results of a usability test.

Another important issue is level of acquaintance. Previous studies have indicated that children behave quite differently depending on how well they know each other. In a study where adult test subjects were asked to bring a friend, co-worker, or family member to the usability test provided a positive experience, i.e. (Kahler, 2000), while other studies stress the importance of using non-acquainted test subjects (Karat et al. 1992). Most studies involving children seems to prefer acquainted pairs of children; this is often achieved through involvement of children attending same school classes or kindergartens (Ellis and Bruckmann, 2001; Montemayor et al., 2002; Rader et al., 1997). In the Eco-I project, the pairing goes beyond acquaintance, since the participating teacher configured the pairs of children according to how well they worked together (Scaife et al., 1997). Few studies indicate that the pairs of children were unacquainted, but this might have been the case in the StoryMat project since the children attended different schools (Cassell and Ryokai, 2001).

When using constructive interaction with children, gender seems to be an important issue as gender does for other aspects of information technology use, e.g. computer games (Gorriz and Medina, 2000) or technology preferences and attitudes (Cassell, 2002; Inkpen, 1997). But it is not apparent if and how gender influences other issues of usability testing, such as effectiveness, efficiency, or number of identified usability problems. Several studies involve both genders in the design processes (Bers et al., 2001; Danesh et al., 2001; Lester et al., 1997; Lumbreras and Sánchez, 1999; Scaife et al., 1997; Stewart et al., 1999; Strommen, 1998). Some studies adapted imbalanced numbers of girls and boys (Benford et al., 2000; Montemayor et al., 2002), while others deliberately chose an equal number of boys and girls (Lester et al., 1997). Furthermore, some studies intentionally use same-sex pairs (Ellis and Bruckmann, 2001; Moher et al., 1999).

Nielsen (1993) claims that constructive interaction is preferable over think-aloud when conducting usability evaluations with children. Where children face difficulties in following the instructions for a think-aloud test, constructive interaction comes closer to their natural behaviour, since the children work in pairs and collaborate in solving the tasks. Due to the fact that the children's ability to verbalize their thoughts and feelings during a test differs, Hanna et al. (1997) propose some adjusted guidelines where they reflect upon common target age ranges. Jensen and Skov (2005) found that 67% of the research on interaction design and children applied some sort of systematic field or laboratory evaluations. Furthermore, some studies have explored different methods for conducting usability evaluations with children; one studied the effectiveness of co-operative evaluations (think-aloud) and

co-discovery evaluations (constructive interaction) (Als et al., 2005, Markopoulos and Bekker, 2003), where another studied different method's effectiveness to elicit verbal comments from children (van Kesteren et al., 2003). The first compared the difference in total number of identifies usability problems identified by four subjects or four pairs, and found only negligible differences between the two methods.

Several studies on usability evaluations with children involve children as subjects applying think-aloud (Druin and Solomon, 1996; Druin, 1999a; Druin, 1999b; Rader et al., 1997; Strommen, 1998), constructive interaction (Moher et al., 1999; Montemayor et al., 2002; Scaife et al., 1997; Skov et al., 2004), or both approaches (Benford et al., 2000; Danesh et al., 2001). However, very few report on how think-aloud or constructive interaction performs as methods for usability problem identification when using children as subjects. Furthermore, none f the studies report on or reflect upon how children should be paired in constructive interaction, e.g. whether to adapt same-sex pairs or having friends act as pairs. This is somewhat surprisingly as previous research on constructive interaction stresses the importance of composing pairs.

3. Method

Our experiment utilized a setup for comparison of think-aloud and constructive interaction for usability testing with children. In particular, we wanted to measure think-aloud and constructive interaction on identification of usability problems and explore the impact of different compositions of pairs in constructive interaction.

Table 1: 60 children participated in our experiment in three different setups: constructive interaction as acquainted dyads or non-acquainted dyads and think-aloud as single testers.

	Single Testers (think-aloud)	Acquainted Dyads (constructive interaction)	Non-Acquainted Dyads (constructive interaction)
Girls	6	6x2	6x2
Boys	6	6x2	6x2
Total	12	12x2	12x2

We designed a between-subject 3x2 experiment with evaluation session setup (single testers, acquainted dyads, non-acquainted dyads) and gender (girls, boys) as independent variable as illustrated in table 1. The primary dependent measures were total number of identified usability problems, average number of usability problems, cost (measured in man hours), unique problems, and subjective workload.

3.1 Test Subjects

60 children (30 girls and 30 boys) at the age of 13 or 14 years old (M=13.35, SD=0.48) participated as test subjects in the experiment. The children were all 7th grade pupils from five different elementary schools in the greater Aalborg area. The children did not receive compensation for their involvement in the experiment.

The children were assigned as test subjects to one of the three test setups e.g. individual testers, acquainted dyads, or non-acquainted dyads. Each setup had twelve individual testers (six girls and six boys), twelve acquainted dyads (six pairs of girls and six pairs of boys), and twelve non-acquainted dyads (six pairs of girls and six pairs of boys). Assignment of the children to the three test setups was done randomly under two conditions. First, all acquainted dyads attended the same school class and secondly all non-acquainted dyads attended different schools. The acquainted pairs had known each

other for at least five years except for one pair of girls and one pair of boys who had been acquainted for one year (*M*=6.25, *SD*=2.5). None of the non-acquainted dyads knew each other in advance.

We assessed all children based on their level of experience with mobile phones. The assessments were made on a scale of 1 to 5 (where 5 was the highest level of expertise). We assessed their expertise from the following five questions: 1) Did they own a mobile phone? 2) How many years had they owned a mobile phone? 3) How many different brands had they previously owned? 4) How many short text messages did they send daily? 5) How many minutes did they talk every day? The questions were each answered from a five-scale by the children themselves individually and all five questions contributed equally to the combined expertise assessment. The mean expertise level for all 60 children was 3.2 (*SD*=1.0) where the girls mean value was 3.2 (*SD*=1.1) and the boys mean value was 3.2 *(SD*=1.0). None of children indicated that they had experience with the mobile phone used in the study.

3.2 System

The selected system for our experiment was an INNO100 mobile phone by Innostream. This particular mobile phone was selected since it had not been released on the European market at the time of our experiment. Thus, all children would have to learn to use the mobile phone.

The inno-100 integrates a range of standard mobile phone features, such as making and receiving phone calls and short text messages, and more advanced features, including speed dial functions and options for creating personalized ring tones. The INNO100 has two separate screens with a main 128x144 pixel 16 bit colour screen and 64x80 pixel sub screen on the cover. The navigation is primarily based on icons in the two upper menu levels. The lower levels are textual based including choice menus for setting values. Furthermore, the INNO100 integrates a number of games.

3.3 Configuring Pairs

As our experiment involved acquainted and non-acquainted dyads, we contacted elementary schools in the greater Aalborg area. A total of five schools agreed to participate and their head teachers authorized us to recruit 7th grade pupils (13-14 years old). Furthermore, we arranged that the tests would take place during normal school hours (from 8.00am to 3.00pm).

Children from the five public schools were introduced to the experiment by two of the participating researchers. These researchers explained their roles in the experiment and how the children's participation would contribute. They were told that they would interact with a mobile phone, but the phone of the experiment was not revealed. Participation in the experiment was voluntarily and interested children got an information sheet describing the experiment in detail and a consent form that had to be signed by a parent or a guardian. After receiving signed consent forms from a total of 60 children, we began scheduling the 36 usability evaluation sessions.

Scheduling the test sessions was done in two steps. First, we matched 48 test subjects for the 12 acquainted dyads and 12 non-acquainted dyads using the following rules 1) acquainted dyads had to attend the same school classes, 2) non-acquainted dyads had to attend different schools, and 3) the 12 acquainted dyads and 12 non-acquainted dyads had to each consist of six pairs of girls and six pairs of boys. The remaining 12 children (6 boys and 6 girls) were assigned as individual testers. Secondly, we arranged the time slots in the children were to act as test subjects.

3.4 Procedure

The sessions were held at the usability laboratory at Aalborg University. We adapted the guidelines for usability testing with children proposed by Hanna et al. (1997). Particularly, we focused on greeting the children, stressing the importance of the participation, and stressing that they were not the object of the test. The purpose of the evaluation was explained in detail to the children and they were shown the facilities of the usability lab. Test subjects intended for roles as non-acquainted dyads were kept separate before the test sessions. The children received questionnaires on which they had to provide answers to a range of questions such as age, name, school, and mobile phone experience. The

usability test sessions were conducted in a specialized usability laboratory. The laboratory integrated two rooms; an observation room in which the evaluations took place and a control room where one of the researchers would handle electronic equipment for recording the sessions and collect data. The two rooms were separated with a one-way mirror allowing people in the control room to see what was going on in the observation room. All sessions were recorded on video for later analysis including facial and body expressions of the children and a close up of the mobile phone to capture their interactions with the mobile phone.

The children were asked to solve twelve tasks one at a time addressing standard and advanced functionalities in the inno-100 mobile phone. This included making a phone call, sending a short text message, adjusting the volume of ring tones, and editing entries in the address book. We did not specify any time limits for the tasks, but required the participants to try to solve all tasks. All children were able to solve all specified tasks. On average, the children spent 26:45 minutes (*SD*=06:39) on the twelve tasks. The individual testers were asked to think-aloud while solving the tasks. We explained think-aloud to the individual testers in terms of the descriptions in Nielsen (1993: pp. 195-198). The acquainted and non-acquainted dyads were asked to solve the tasks using constructive interaction where they should collaborate in order to solve the assigned tasks. We explained constructive interaction to the dyads in terms of the descriptions in Nielsen (1993: p. 198).

After the usability sessions, the children completed a subjective workload test NASA-TLX (Miller and Hart, 1984). The children filled in the tests individually even when participating in pairs. This was done to evaluate the workload as experienced by the children in order to compare the different setups. We translated the test into the children's native language, Danish.

3.5 Data Analysis

Two researchers conducted all 36 evaluations, acting as test monitor and logger as defined in (Rubin, 1994). The two researchers analyzed all of the video recorded from the usability sessions. A collaborative approach was used to discuss and classify usability problems. Problems were classified according to severity using the instrumentation in Molich (1997) extended with classification of serious and cosmetic problems. Problem severity classification was discussed until consensus was reached. Sessions were picked randomly for analysis in order to avoid biasing the rating process. The NASA-TLX tests were further analyzed. 55 tests were answered correctly by the children while 5 were answered incompletely.

We further calculated the proportions of problems identified with different numbers of subjects and sessions as illustrated in (Bekker et al., 2008; Nielsen and Landauer, 1993). Increased numbers of involved subjects usually generates a higher number of identified usability problems. Nielsen and Landauer (1993) display this correlation between subject numbers and problem numbers and propose a mathematical proportions model for identifying usability problems based on number of subjects or evaluators. Proportions of usability problems can be estimated through the Poisson model (Nielsen and Landauer, 1993; in our experiment exact values for every session and subject number were calculated. Through combinations of increased numbers of sessions for the three setups, we show the calculated total number of identified problems with increased numbers of sessions ranging from one session to 12 sessions. The number of combinations of k objects from a set of n objects was calculated from

$$\text{(EQ1):} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where k is numbers between 1 and 12 and n=12. As an example, calculating number of problems for the three sessions (k=3, n=12) involves 220 different combinations (according to EQ1). For all 220 combinations, we counted the number of identified problems and afterwards calculate the mean number for all 220 combinations.

Furthermore, as a result of the calculated proportions of identified usability problems, we calculated cost/benefit ratio for all numbers of subjects or sessions. Defining cost for this experiment, we use the number of evaluator hours spent on setting up the test, conducting the test, and analyzing the

test. Thus, the number of hours delivered by the children does not influence our cost. The cost of setting up and designing the usability test with the children for our experiment was constant regardless the number of sessions. The cost/benefit-ratio is calculated from

$$\text{(EQ 2):} \qquad Ratio(n) = \frac{(C + (x + y) * 2 * n)}{Found(n)}$$

where n is number of sessions (subjects), Found (n) is the number of problems found for n sessions (subjects), C is the overhead costs of setting up the entire test, x is the mean task completion time for the actual setup, and y is mean analysis time for analyzing the results for the actual setup. Mean task completion times and mean analysis times are multiplied with 2 as two evaluators participated in the conduction and analysis.

We applied different statistical analyses on the results. We used one-way analysis of variance tests (ANOVA) with individual testers, acquainted dyads, and non-acquainted as independent variables and with Tukey HSD post-hoc tests. For comparison of two means (primarily comparison of gender), we used Mann-Whitney tests.

4. Results

The 60 children in the 36 usability test sessions solved all 12 assigned tasks. Even though the constructive interaction sessions with acquainted dyads (*M*=29:54, *SD*=06:57) spent most time on the assignments in our experiment; the individual testers (*M*=25:34, *SD*=03:44), and the non-acquainted dyads (*M*=24:48, *SD*=07:53), we found no significant differences for the task completion times.

4.1 Total Numbers of Identified Usability Problems

The results from the 36 usability test sessions resulted in the identification of 81 different usability problems (see table 2). Based on our classification scheme, we classified 32 of these 81 usability problems as critical problems, 13 as serious problems, and 36 as cosmetic problems.

Table 2. Key results from the experiment on problem identification. The table illustrates the number of identified critical, serious and cosmetic problems (plus the sum of the three severities) in the three setups: individual testers, acquainted dyads, and non-acquainted dyads.

	Individual Testers (N=12)	Acquainted Dyads (N=12)	Non-Acquainted Dyads (N=12)	Total (N=36)
Critical	25	28	22	32
Serious	8	13	6	13
Cosmetic	23	25	23	36
Sum	56	66	51	81

Our experiment exposed differences in problem identification between think-aloud and constructive interaction where we found that the constructive interaction sessions with acquainted dyads identified 18% more different problems than the think-aloud sessions. The acquainted dyads identified the highest number of usability problems of the three setups with a total of 66 of the 81 identified usability problems (81%) while the individual testers identified 56 of the 81 usability problems (69%). The non-acquainted dyads identified only 51 of the 81 usability problems (63%).

Looking at problem severity, we further found that the acquainted dyad sessions identified nearly all critical problems namely 28 of the 32 (88%), whereas the individual testers identified 25 of the 32 problems (78%) and the non-acquainted dyads experienced 22 of the 32 problems (69%). We found a similar pattern for the serious problems with acquainted dyad sessions identifying 12 of 13 problems

(92%), individual testers 8 of the 12 problems (67%), non-acquainted dyads 6 of the 12 problems (50%). Thus, regarding the most severe identified problems, the acquainted dyad sessions again facilitated the identification of the highest number of usability problems.

4.2 Average Numbers of Identified Problems

The sessions exhibited great variance in number of identified problems. As an example, one acquainted girl session facilitated the identification of 22 different usability problems while another acquainted girl session only facilitated 10 different problems. This pattern was discovered throughout all the sessions in the three setups.

Analyzing the average numbers of identified problems, we found visible deviations between the setups; acquainted dyads identified 17.17 problems (*SD*=5.06), non-acquainted dyads identified 15.08 problems (*SD*=4.87), and individual testers identified 13.50 problems (*SD*=2.24). The somewhat high standard deviations indicate great variances between the setups and we found no significant differences between the three setups according to a one-way ANOVA test $F_{(2,33)}=2.150$, $p=0.133$. Furthermore, we found no significant differences for neither the critical problems $F_{(5,30)}=1.875$, $p=0.128$, nor for the identified serious problems $F_{(5,30)}=1.320$, $p=0.282$, or for the identified cosmetic problems $F_{(5,30)}=1.050$, $p=0.407$.

We further calculated proportions of problems identified with different numbers of subjects and sessions (as performed in Bekker et al., 2008). The graph in figure 1 illustrates a clear logarithmic distribution for the three setups where the acquainted dyads for all calculated numbers of sessions identified higher numbers of usability problems. The figure clearly illustrates that acquainted dyads identify more usability problems for all session numbers. For any two sessions, we calculated that the acquainted dyads would identify 28.52 problems on average while non-acquainted dyads would identify 25.08 problems on average and think-aloud subjects 22.89 on average and this difference is significant according to a one-way ANOVA test $F_{(2,195)}=30.677$, $p=0.0001$. A post-hoc Tukey HSD comparison showed significant different at the 1% level between the acquainted dyads and both single testers and non-acquainted dyads as well as significant difference at the 1% level between the non-acquainted dyads and single testers.

As constructive interaction by nature involves twice as many test subjects per session as think-aloud, we further calculated numbers of problems identified with increasing numbers of subjects (see figure 1, right). Perhaps not surprisingly, we found that think-aloud per subject identified more problems than both constructive interaction setups. As an example, involving six subjects in a usability test (a typical number of subjects for many usability studies), the think-aloud would potentially generate 41.34 usability problems (the same as in figure 1, left as sessions numbers are equal to subject numbers for think-aloud), but with six subjects constructive interaction with acquainted dyads would generate 35.91 usability problems, whereas the constructive interaction with non-acquainted dyads would generate 31.22 usability problems. From only two subjects, the think-aloud protocol produced significantly higher numbers of problems than both the two constructive interaction setups – for two subjects $F_{(2,87)}=33.328$, $p=0.0001$.

Even though we identified no significant differences between average numbers of identified usability problems, with increased numbers of sessions or subjects we discovered differences between the setups. If access to children subjects is difficult or numbers are scarce, our results indicated that in terms of problem identification evaluators should consider using think-aloud with individual testers. However, if subjects are not scarce constructive interaction with acquainted dyads seemed preferable.

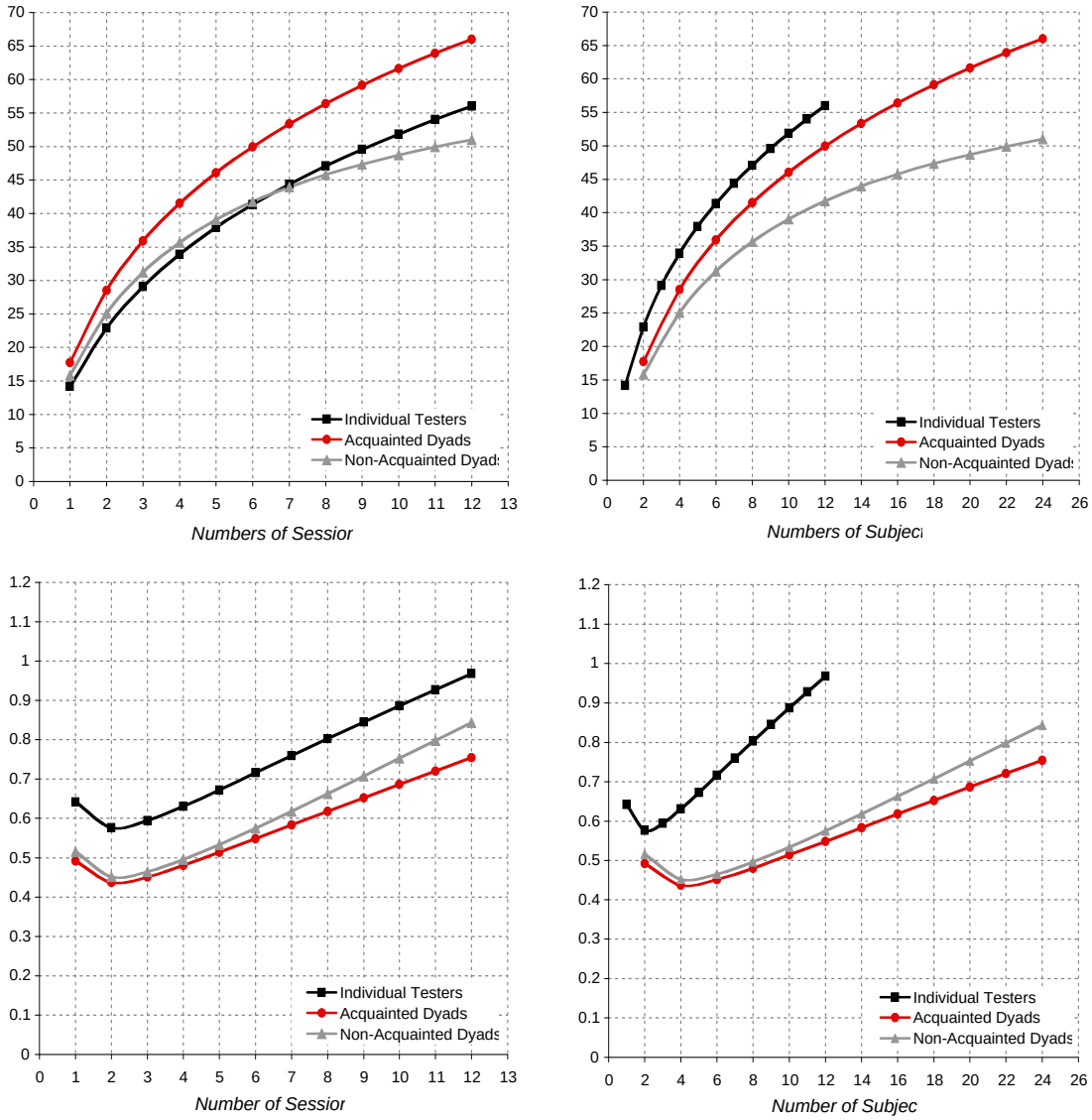

Figure 1: Calculated numbers of identified problems per numbers of sessions and numbers of subjects (top). For all numbers of sessions, constructive interaction with acquainted dyads identified a larger amount of problems (top, left). Further, for any number of involved subjects individual testers using think-aloud identified a larger number of problems (top, right). At the bottom the figure, we calculated ratio between benefits (numbers of identified usability problems) and costs (evaluator hours) with increasing numbers of sessions and subjects for the three setups. Cost is calculated from evaluator hours spent on setting up the test, conducting the test, and analyzing the test

4.3Cost against Benefits

Having calculated proportions of problems identified with increasing numbers of sessions and subjects, we analyzed the three setups in relation to their potential cost and benefits. We define benefits as the total number of problems identified by any number of sessions or subjects.

Figure 1 (bottom) shows the cost/benefit-ratio (number of evaluator hours /number of identified problems) for increasing numbers of sessions. As indicated by the figure, the two constructive interaction setups follow the same pattern and are very close with a marginal advantage to the acquainted dyads. Reaching the lowest cost just after two sessions, the constructive interaction with acquainted dyads had an estimated cost of 0.44 hours (~26 min) per usability problem whereas constructive inter-

action with non-acquainted dyads had a cost of 0.45 hours (~27 min) per usability problems, and think-aloud was 0.58 hours (~35 min) per usability problem and this difference is significant $F_{(2,195)}=49.829$, p=0.0001. A post-hoc test revealed a significant difference between at the 1% level between both the acquainted dyads and individual testers and non-acquainted dyads and individual testers. On the other hand, we found no significant difference in cost/benefit for involving only one subject.

Again as constructive interaction inherently involves twice as many subjects per session as think-aloud, considering the cost/benefit-ratio, think-aloud costs twice as much as constructive interaction per subject in the conduction and analysis. Therefore, we also calculated the ratio between benefits and costs with increasing numbers of subjects (figure 2, right). This figure illustrates that having individual testers in an evaluation clearly heightens the cost for the participating evaluators compared to dyads. The additional test conductions and rounds of analyses make the cost/benefit-ratio less attractive.

4.4 Identification of Unique Problems

Different usability evaluation methods often uncover unique problems only discovered by certain kinds of methods. In our experiment, we distinguished between two kinds of unique problems inspired by Karat et al. (1992). Firstly, we summarized problems that were identified in one session only (similar to action areas in Karat et al. (1992)). Secondly, we summarized problems identified only in one of the three setups but by at least two sessions (similar to unique usability problem areas in Karat et al. (1992)).

	Individual Testers (N=12)	Acquainted Dyads (N=12)	Non-Acquainted Dyads (N=12)	Sum (N=36)
Critical	2 (2)	4 (0)	0 (0)	6 (2)
Serious	1 (0)	2 (2)	0 (0)	3 (2)
Cosmetic	6 (0)	6 (0)	3 (0)	15 (0)
Sum	9 (2)	12 (2)	3 (0)	24 (4)

Table 3: Identification of unique problems. The numbers outside parentheses signify the numbers of unique problems identified by exactly one session (No action areas in Karat et al. (1992)). Numbers in parentheses denote the numbers of problems identified in only one setup but in at least two sessions

As table 3 shows, six of the 36 critical problems (17%), three of the 13 serious problems (23%), and 15 of the 36 cosmetic problems (42%) were identified in one session only (no action areas). Thus, almost half of the cosmetic problems were identified in only one of the 36 usability sessions.

Think-aloud with individual testers identified some critical problems not identified by constructive interaction sessions. Specifically, think-aloud identified four unique critical problems of which two were identified in at least two sessions. Finally, the identified unique problems showed that constructive interaction with non-acquainted dyads identified no critical or serious problem that was not identified by either both think-aloud or constructive interaction with acquainted dyads.

5. Discussion

Our experiment was partly inspired by Nielsen (1993) who claims that that constructive interaction is preferable over think-aloud when conducting usability evaluations with children as children often face difficulties in following the instructions for a think-aloud test. Thus, constructive interaction comes closer to their natural behaviour since the children work in pairs and collaborate in solving the tasks. Much research has been conducted with the involvement of children as subjects applying think-aloud (Druin and Solomon, 1996; Druin, 1999a; Druin, 1999b; Rader et al., 1997; Strommen, 1998), con-

structive interaction (Moher et al., 1999; Montemayor et al., 2002; Scaife et al., 1997; Skov et al., 2004), or both approaches (Benford et al., 2000; Danesh et al., 2001). Markopoulos and Bekker (2003) investigated usability testing with children on think-aloud (co-operative evaluation) and constructive interaction (co-discovery evaluation) and they found only small differences between the two methods on problem identification. The act of verbalizing often makes a significant basis for the identification and classification of usability problems in subsequent data analyses. Based on Nielsen's assumption, constructive interaction could lead to the identification of a higher number of usability problems when testing with children. Our experiment and results seemed to confirm this, at least partially.

Our results illustrated significant differences between classical think-aloud and constructive interaction when usability testing with children. Further, we found that the composition of pairs in constructive interaction had significant effects on the identification of usability problems where acquainted dyads identified a higher number of problems compared to the non-acquainted dyads.

Nielsen (1993) stresses the aptness of constructive interaction in projects where access to large number of test subjects is easy. For certain kinds of studies, limited access to test subjects is inevitable, e.g. if it is a requirement that the participating children suffer from a specific decease, as seen in [2, 3]. Our study partially supports this, as 4 individual testers are likely to identify more problems than 2 pairs in constructive interaction. Our study stressed the importance of pairing subjects in acquainted dyads as they identified a higher number of problems. Other studies have paired the children according to how well a teacher believes they would work together (Scaife et al., 1997). However, as we paired the children randomly under the conditions of acquaintance and non-acquaintance we have no immediate results supporting or rejecting this issue. Markopoulos and Bekker (2993) provide no answer to their configuration of the pairs in their study.

Hanna et al. (1997) argue that children are cognitively diverse when involved in usability evaluations. A primary element is children's age which highly influences their abilities to take active part in the test situation. Hanna et al. state that many 13-14 years old children will be able to think-aloud; while others may be too self-conscious about having people watching them. We experienced no major problems related thinking-aloud for the children working alone. However, their verbalization facilitated a lower number of identified problems.

6. Conclusion

Constructive interaction has been suggested as a suitable usability evaluation method when usability testing with children. It may be difficult to get children to follow the instructions for a standard thinking-aloud test. We presented an experiment that compared think-aloud and constructive interaction in usability testing with 13-14 years old children with a special focus on how pairs of children should be configured in constructive interaction.

Constructive interaction with pairs of children knowing each other identified more problems (on all severities) and specifically more critical problems. We also found that the children age 13-14 years old had no major problems in following the standard thinking-aloud protocol. Furthermore, we found that the composition of pairs had impact on the problem identification. Acquainted dyads identified a higher number of usability problems compared to non-acquainted dyads. Especially for the calculated proportions of identified usability problems, we found that acquainted dyads of children identified experienced more problems than both single testers (using think-aloud) and non-acquainted dyads (using constructive interaction). Finally, acquainted dyads identified more unique problems than any of the two other conditions.

Our study suffers from a number of limitations, which could form further research with children. First, our results of our experiment cannot simply be generalized for all ages of children. Thus, replicating the experiment with younger children may show a different kind of relationship between think-aloud and constructive interaction. Secondly, we recorded that the non-acquainted dyads continuously took turns with the mobile phone making it difficult for the other child to see what was going on at the interface. This could probably be different for desktop-based applications. Future usability testing with

children should consider which usability evaluation method to adapt. They should carefully consider the pair configuration when choosing constructive interaction. Based on the access to children as test subjects, they should consider choosing think-aloud if access to children is limited.

Acknowledgements

The work behind this paper received financial support from the Danish Research Agency (grant no. 2106-04-0022). We would especially like to thank all the participating children and their parents. Furthermore, we would like to thank Kori Inkpen for comments and several anonymous reviewers for comments on drafts of this paper.

References

Als, B. S., Jensen, J. J., and Skov, M. B. (2005) Comparison of Think-Aloud and Constructive Interaction in Usability Testing with Children. In *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, ACM Press

Bekker, M. M., Baauw, E., and Barendregt, W. (2008) A comparison of two analytical evaluation methods for educational computer games for young children. Cognition, Technology, and Work. DOI 10.1007/s10111-007-0068-x

Benford, S., Bederson, B. B., Åkesson, K-P, Bayon, V., Druin, A., Hansson, P., Hourcade, J. P., Ingram, R., Neale, H., O'Malley, C., Simsarian, K. T., Stanton, D., Sundblad, Y., and Taxén, G. (2000) Designing storytelling technologies to encouraging collaboration between young children. In *Proceedings of the Human Factors and Computing Systems CHI'00*, ACM Press, pp. 556 - 563

Bers, M. U., Gonzalez-Heydrich, J., and DeMaso, D. R. (2001) Identity Construction Environments: Supporting a Virtual Therapeutic Community of Pediatric Patients Undergoing Dialysis. In *Proceedings of the Human Factors and Computing Systems CHI'01*, ACM Press, pp. 380 - 387

Cassell, J. (2002) *Genderizing*. The Handbook of Human-Computer Interaction

Cassell, J. and Ryokai, K. (2001) Making Space for Voice: Technologies for Supporting Children's Fantasy and Storytelling. *Personal and Ubiquitous Computing*, Springer-Verlag, vol. 5(3), pp. 169 - 190

Danesh, A., Inkpen, K. M., Lau, F., Shu, K., Booth, K. S. (2001) Geney: Designing a collaborative activity for the Palm handheld computer. In Proceedings of the Human Factors and Computing Systems CHI'01, ACM Press, pp. 388 - 395

Druin, A. and Solomon, C. (1996) Designing Multimedia Environments for Children. Wiley & Sons, New York

Druin, A. (1999a) The Role of Children in the Design of New Technology. HCIL Technical Report No. 99-23, University of Maryland, USA

Druin, A. (1999b) The Design of Children's Technology. Morgan Kaufmann Publishers, Inc., San Francisco, CA

Ellis, J. B. and Bruckman, A. S. (2001) Designing Palaver Tree Online: Supporting Social Roles in a Community of Oral History. In Proceedings of the Human Factors and Computing Systems CHI'01, ACM Press, pp. 474 - 481

Ericsson, K.A. and Simon, H.A. (1990) *Protocol Analysis. Verbal reports as data*, Cambridge Massachusetts

Gorriz, C. M. and Medina, C. (2000) Engaging Girls with Computers through Software Games. Communications of the ACM, vol. 43, No. 1, pp. 42 – 49

Gray, W. D. and Salzman, M. C. (1998) Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. In Human-Computer Interaction, vol. 13, pp. 203 – 261

van den Haak, M. J., de Jong, and Schellens, P. J. (2004). Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. Interacting with Computers, 16(6):1153-1170

Hanna, L., Risden, K., and Alexander, K. J. (1997) Guidelines for Usability Testing with Children. In interactions, September + October, pp. 9 – 14

Inkpen, K. (1997) Three Important Research Agendas for Educational Multimedia: Learning, Children, and Gender. In Proceedings of Educational MultiMedia '97

Jensen, J. J. and Skov, M. B. (2005) A Review of Research Methods in Children's Technology Design. In *Proceedings of the 4th International Conference on Interaction Design and Children (IDC'05)*, ACM Press

Kahler, H. (2000) Constructive Interaction and Collaborative Work. interactions, May + June, pp. 27 - 34

Karat, C.-M., Campbell, R., and Fiegel, T. (1992) Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. In Proceedings of the Human Factors and Computing Systems CHI'92, ACM Press, pp. 397-404

van Kesteren, I. E. H., Bekker, M. M., Vermeeren, A. P. O. S., and Lloyd, P. A. (2003) Assessing usability evaluation methods on their effectiveness to elicit verbal comments from children subjects. In Proceeding of the 2003 conference on Interaction design and children (IDC'03), ACM Press, pp. 41 - 49

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., and Bhogal, R. S. (1997) The Persona Effect: Affective Impact of Animated Pedagogical Agents. In Proceedings of the Human Factors and Computing Systems CHI'97, ACM Press, pp. 359 - 366

Lumbreras, M. and Sánchez, J. (1999) Interactive 3D Sound Hyperstories for Blind Children. In Proceedings of the Human Factors and Computing Systems CHI'99, ACM Press, pp. 318 - 325

Markopoulos, P. and Bekker, M. (2003) On the Assessment pf Usability Testing Methods for Children. Interacting with Computers, Elsevier, Vol. 15, pp. 227 – 243

Miller R. C. and Hart, S. G. (1984) Assessing the Subjective Workload of Directional Orientation Tasks. In Proceedings of 20th Annual Conference on Manual Control, NASA Conference Publication, pp. 85 – 95

Miyake, N. (1986) Constructive Interaction and the Iterative Process of Understanding. Cognitive Science, vol. 10(2), pp. 151 - 177

Moher, T., Johnson, A., Ohlsson, S., and Gillingham, M. (1999) Bridging Strategies for VR-based Learning. In Proceedings of the Human Factors and Computing Systems CHI'99, ACM, pp. 536 - 543

Montemayor, J., Druin, A., Farber, A., Simms, S., Churaman, W., and D'Amour, A. (2002) Physical Programming: Designing Tools for Children to Create Physical Interactive Environments. In Proceedings of the Human Factors and Computing Systems CHI'02, ACM Press, pp. 299 - 306

Nielsen, J. (1993) Usability Engineering. Academic Press

Nielsen, J. and Landauer, T. K. (1993) A Mathematical Model of the Finding of Usability Problems. In Proceedings of the Human Factors and Computing Systems INTERCHI'93, ACM Press, pp. 206 - 213

O'Malley, C. E., Draper, S. W., and Riley, M. S. (1984) Constructive Interaction: A Method for Studying Human-Computer-Human Interaction. In Proceedings of IFIP Interact '84, pp. 269 – 274

Rader, C., Brand, C., and Lewis, C. (1997) Degrees of Comprehension: Children's Understanding of a Visual Programming Environment. In Proceedings of the Human Factors and Computing Systems CHI'97, ACM Press, pp. 351 - 358

Resnick, M., Martin, F., Berg, R., Borovoy, R., Colella, V., Kramer, K., and Silverman, B. (1998) Digital Manipulatives: New Toys to Think With. In Proceedings of the Human Factors and Computing Systems CHI'98, ACM, pp. 281 – 287

Scaife, M., Rogers, Y., Aldrich, F., and Davies, M. (1997) Designing for or Designing with? Informant Design for Interactive Learning Environments. In Proceedings of the Human Factors and Computing Systems CHI'97, ACM Press, pp. 343 - 350

Skov, M. B., Andersen, B. L., Duhn, K., Garnæs, K. N., Grünberger, O., Kold, U., Mortensen, A. B., and Sørensen, J. A. L. (2004) Designing a Drawing Tool for Children: Supporting Social Interaction and Communication. In Proceedings of the Australian Computer-Human Interaction Conference 2004 (OzCHI'04)

Stewart, J., Bederson, B. B., and Druin, A. (1999) Single Display Groupware: A Model for Co-Present Collaboration. In Proceedings of the Human Factors and Computing Systems CHI'99, ACM, pp. 286 - 293

Strommen, E. (1998) When the Interface is a Talking Dinosaur: Learning across Media with ActiMates Barney. In Proceedings of the Human Factors and Computing Systems CHI'98, ACM Press, pp. 288 – 295

		Number of sessions											
		1	2	3	4	5	6	7	8	9	10	11	12
Individual testers	N=12	14.17 (2.11)	22.89 (2.28)	29.10 (2.36)	33.93 (2.41)	37.92 (2.39)	41.34 (2.32)	44.36 (2.21)	47.08 (2.06)	49.57 (1.86)	51.86 (1.58)	54.00 (1.15)	56.00 (0.00)
Acquainted Dyads	N=12	17.75 (5.05)	28.52 (5.05)	35.91 (4.90)	41.52 (4.78)	46.08 (4.64)	49.95 (4.43)	53.35 (4.16)	56.39 (3.80)	59.14 (3.35)	61.64 (2.77)	63.92 (1.98)	66.00 (0.00)
Non-Acquainted Dyads	N=12	15.83 (4.41)	25.08 (4.51)	31.22 (4.14)	35.66 (3.73)	39.05 (3.33)	41.74 (2.94)	43.94 (2.58)	45.78 (2.23)	47.35 (1.90)	48.71 (1.54)	49.92 (1.11)	51.00 (0.00)
Individual Girls	N=6	14.67 (2.43)	22.87 (2.60)	28.15 (2.20)	32.00 (1.75)	35.17 (1.34)	38.00 (0.00)	--	--	--	--	--	--
Individual Boys	N=6	13.67 (1.60)	22.73 (1.65)	29.55 (1.69)	35.20 (1.60)	40.00 (1.15)	44.00 (0.00)	--	--	--	--	--	--
Acquainted Girl Dyads	N=6	18.17 (5.05)	29.47 (5.06)	37.40 (4.64)	43.60 (3.96)	48.67 (2.98)	53.00 (0.00)	--	--	--	--	--	--
Acquainted Boys	N=6	17.33 (5.02)	28.00 (4.49)	35.00 (3.54)	40.00 (2.73)	43.83 (1.95)	47.00 (0.00)	--	--	--	--	--	--
Non-Acquainted Girls	N=6	18.00 (4.36)	28.60 (3.84)	35.55 (2.85)	40.67 (2.05)	44.67 (1.37)	48.00 (0.00)	--	--	--	--	--	--
Non-Acquainted Boys	N=6	13.67 (3.24)	21.73 (2.93)	27.35 (2.73)	31.53 (2.22)	34.67 (1.49)	37.00 (0.00)	--	--	--	--	--	--
All Girls	N=18	16.94 (4.40)	27.01 (4.74)	33.88 (4.73)	39.03 (4.66)	43.16 (4.58)	46.62 (4.49)	49.61 (4.40)	52.27 (4.31)	54.67 (4.20)	56.88 (4.09)	58.95 (3.94)	60.89 (3.77)
All Boys	N=18	14.89 (3.97)	24.08 (4.26)	30.42 (4.24)	35.19 (4.13)	39.02 (3.97)	42.22 (3.79)	44.97 (3.62)	47.38 (3.44)	49.51 (3.25)	51.43 (3.06)	53.16 (2.86)	54.75 (2.65)

Appendix A: Number of identified problems for any subject number calculated from all combinations

		Number of subjects																						
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Individual testers	N=12	14.17 (2.11)	22.89 (2.28)	29.10 (2.36)	33.93 (2.41)	37.92 (2.39)	41.34 (2.32)	44.36 (2.21)	47.08 (2.06)	49.57 (1.86)	51.86 (1.58)	54.00 (1.15)	56.00 (0.00)	--	--	--	--	--	--	--	--	--	--	--	--
Acquainted Dyads	N=12	--	17.75 (5.05)	--	28.52 (5.05)	--	35.91 (4.90)	--	41.52 (4.78)	--	46.08 (4.64)	--	49.95 (4.43)	--	53.35 (4.16)	--	56.39 (3.80)	--	59.14 (3.35)	--	61.64 (2.77)	--	63.92 (1.98)	--	66.00 (0.00)
Non-Acquainted Dyads	N=12	--	15.83 (4.41)	--	25.08 (4.51)	--	31.22 (4.14)	--	35.66 (3.73)	--	39.05 (3.33)	--	41.74 (2.94)	--	43.94 (2.58)	--	45.78 (2.23)	--	47.35 (1.90)	--	48.71 (1.54)	--	49.92 (1.11)	--	51.00 (0.00)
Individual Girls	N=6	14.67 (2.43)	22.87 (2.60)	28.15 (2.20)	32.00 (1.75)	35.17 (1.34)	38.00 (0.00)	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Individual Boys	N=6	13.67 (1.60)	22.73 (1.65)	29.55 (1.69)	35.20 (1.60)	40.00 (1.15)	44.00 (0.00)	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Acquainted Girl Dyads	N=6	--	18.17 (5.05)	--	29.47 (5.06)	--	37.40 (4.64)	--	43.60 (3.96)	--	48.67 (2.98)	--	53.00 (0.00)	--	--	--	--	--	--	--	--	--	--	--	--
Acquainted Boys	N=6	--	17.33 (5.02)	--	28.00 (4.49)	--	35.00 (3.54)	--	40.00 (2.73)	--	43.83 (1.95)	--	47.00 (0.00)	--	--	--	--	--	--	--	--	--	--	--	--
Non-Acquainted Girls	N=6	--	18.00 (4.36)	--	28.60 (3.84)	--	35.55 (2.85)	--	40.67 (2.05)	--	44.67 (1.37)	--	48.00 (0.00)	--	--	--	--	--	--	--	--	--	--	--	--
Non-Acquainted Boys	N=6	--	13.67 (3.2	--	21.73 (2.9	--	27.35 (2.7	--	31.53 (2.2	--	34.67 (1.4	--	37.00 (0.0	--	--	--	--	--	--	--	--	--	--	--	--

			4)		3)		3)		2)		9)		0)													
All Girls	N=18	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
All Boys	N=18	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Evaluating in a Healthcare Setting: A Comparison between Concurrent and Retrospective Verbalisation

Janne Jul Jensen

Department of Computer Science, Aalborg University
Fredrik Bajers Vej 7, E2-220, DK-9210 Aalborg East, Denmark
jjj@cs.aau.dk

Abstract. The think-aloud protocol, also known as concurrent verbalisation protocol, is widely used in the field of HCI today, but as the technology and applications have evolved the protocol has had to cope with this. Therefore new variations of the protocol have seen the light of day. One example is retrospective verbalisation. To compare concurrent and retrospective verbalisation an experiment was conducted. A home healthcare application was evaluated with 15 participants using both protocols. The results of the experiment show that the two protocols have each their strengths and weaknesses, and as such are very equally good although very different.

Introduction

One of the most commonly used protocols in usability evaluations is think-aloud. It is also known under the name concurrent verbalisation, which will be the term used in this paper. Concurrent verbalisation was originally introduced by Karl Duncker [2] and has since then incorporated into HCI. Some of the strengths of the protocol are that it is easy to learn [1, 5], it can be used by non-specialists in usability [6] and it gives a fairly good insight into the cognitive processes of the participant in the evaluation [4]. However, over the years some weaknesses have also been revealed. These include a heightened mental workload of the participant [8] and that the thinking aloud disturbs the participant's interaction with the application [7].

Originally in HCI concurrent verbalisation was used in laboratory settings, but as applications have evolved and become both mobile and context aware among other things, the protocol has been challenged to cope with these new changes. Similarly to bringing telephone conversations out into the public space, using think-aloud in all settings might prove troublesome. Take, for instance, a newer branch of applications for families or friends. Here we are dealing with information that can be very private to the people involved and thus a certain amount of awkwardness can be expected if they are to verbalise this in an evaluation.

If verbalisation in the classical sense of concurrent verbalisation is not always appropriate, then it is necessary to think in alternatives. Another version of verbalisation that has been used in several contexts is retrospective verbalisation. Just like concurrent verbalisation this protocol has both strengths and weaknesses. One advantage is

the decrease of mental workload, as the participant is now free to focus on the task at hand. However, a drawback could be that participants quickly forget specific details that occurred in the task solving process and they are then unable to recall these details afterwards [3]. To shed some light on the pros and cons of the two protocols an experiment was conducted. This was done as a field evaluation in the home healthcare system. The reason for choosing this setting and type of evaluation was to make the setting as realistic as possible in order to investigate any possible effects the surroundings might have with regards to sensitivity. Is it possible to observe any awkwardness in using the concurrent think-aloud protocol compared to the retrospective think-aloud protocol, with respect to a sensitive setting?

The Experiment

To compare concurrent vs. retrospective verbalisation in a healthcare setting and to test the appropriateness of each protocol, an experiment was conducted. It was set up as a field evaluation to create as realistic settings as possible.

The system chosen for evaluation was an application developed to aid home healthcare workers in their daily work. It is an electronic replacement to the existing paper-based system which is currently in use in many municipalities in Denmark. It supports the current work-procedure as well as offer new functionality such as wireless access to added information about the elder citizens and the progress of co-workers, information that earlier was available only at the main office building.

Participants

15 participants were chosen with the help of the head of the group of home healthcare workers with due consideration for work plans etc. All 15 were trained home healthcare workers and their demographic data is shown in table 1.

Table 1. The demographic data of the 15 participants in the two protocols.

Protocol		Age	Experience local	Experience total	Experience computer (1-6)
Retrospective	Average	42.0	5½	8 ¼	3
	High	54	12	13	6
	Low	33	2½	3 ¾	1
Concurrent	Average	42.4	7	10.3	3.9
	High	57	18	23	6
	Low	31	1	1½	1

The table shows the age, the experience as home healthcare workers in the municipality where the experiment took place, the experience as home healthcare workers in

total and the level of experience with computers on a scale from one to six where 1 is most experienced and 6 is least experienced. For each of these variables, the high low and average has been calculated for each of the protocols.

Equipment

Fig. 1. The small clip-on wireless mobile camera from the mobile laboratory.

Fig. 2. The equipment in the mobile laboratory used for concurrent verbalisation.

Fig. 3. The mobile laboratory packed up for use.

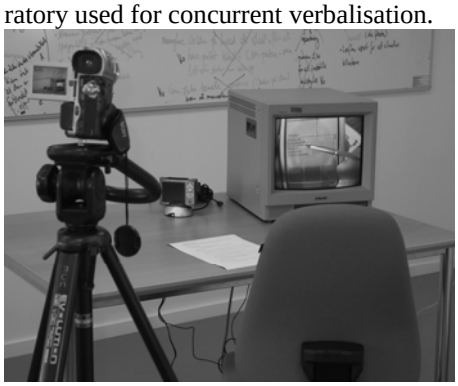

Fig. 4. The setup for retrospective verbalisation.

To support the field evaluation a mobile laboratory was used. It consists of small clip-on wireless mobile cameras (see figure 1), wireless microphones and a mobile digital video recorder. To run it all, it furthermore requires various types of batteries and receivers for the wireless technology. Only the camera and microphone are carried by the participant, the rest is carried by the test monitor packed in a small bag (see figure 2 and 3).

For retrospective verbalisation, the digital recordings from the mobile video recorder were played back to the participant and the retrospective verbalisation was caught using a camcorder (see figure 4).

Procedure

To gain the necessary insight into the field of home healthcare, a small ethnographic field study was conducted. Based on a thorough examination of the system and the insight gained from the ethnographic field study the 8 tasks that covered a wide range of the commonly used functionalities in the application were designed and the experiment was then designed in detail. With the design of the experiment in place, a pilot was conducted for both protocols and the setup was adapted according to the minor issues discovered.

15 participants were recruited from a local municipality. 14 were female and one male, which was representative for the employment situation where women far outweighed the men. The actual experiment took six days and all evaluations were recorded on video. The evaluations took place in six different homes of actual elderly citizens, with the citizen present during the evaluation to further heighten the realism in the experiment.

7 of the 15 participants were assigned to evaluate using retrospective verbalisation, while the remaining 8 participants evaluated the application using concurrent verbalisation. Each of the participants was given a thorough introduction to the experiment, explaining the equipment and its function, what their contribution was, what was expected of them, what would happen etc. They were also instructed thoroughly in how to apply the protocol assigned to them. They were then given 10 minutes to freely familiarise themselves with the system, before trying to solve the tasks.

After the introduction the experiment itself took place in the home of an elderly citizen where the participants attempted to solve the tasks handed out. 8 participants solved them thinking aloud during the evaluation whereas the other 7 had their test session played back to them on a screen afterwards and were thinking aloud during the replay. Upon completion of the evaluation each participant was debriefed.

All the raw video data was analysed afterwards and a list of problems was constructed. The severity of each of the problems was categorised according to the definition by Rolf Molich [5]. According to the definition a problem experienced by a participant falls in one of three categories:

- **Cosmetic:** The user is delayed for less than one minute, is mildly irritated or is confronted with information, which to a lesser degree deviates from the expected.
- **Serious:** The user is delayed for several minutes, is somewhat irritated or is confronted with information, which to some degree deviates from the expected.
- **Critical:** The users attempt to solve the task comes to a halt; the user is very irritated or is confronted with information which to a critical degree deviates from the expected.

The categorisation was done by observing the video recording of each participant, and then evaluate each situation according to the guidelines described above. A given problem is often not experienced equally serious by each participant, and in those cases the problem is categorised in the most severe category.

Results

This section sums up the observations made from the list of problems, which was extracted from the analysis of the raw video data.

Problems Revealed

In total, 105 problems were identified through the evaluation and interestingly the participants using concurrent verbalisation revealed a total of 87 problems whereas the participants using retrospective verbalisation only experienced 61 problems in total. This is a quite big difference which origin is not clear. One explanation could be that the participants evaluating with retrospective verbalisation has an average computer experience level that is almost a point better (3.0) compared to that of the participants using concurrent verbalisation (3.9) on a scale from 1 to 6 (see table 2).

Table 2. Total number of problems, unique problems and the average computer skill of the participants.

	All	Concurrent Verbalisation	Retrospective Verbalisation
Problems revealed	105	87	61
Unique problems*	44	30 (47)	14 (33)
Average computer experience	3.4	3.9	3.0

* Note that the number in parentheses refers to problems that are unique to that protocol and not necessarily unique in total.

Unique Problems

When looking at the number of unique problems the experiment in total reveals 44 unique problems. 30 of these are problems revealed by the concurrent verbalisation protocol, whereas the retrospective verbalisation protocol only experience 14 of the 44. Even if we look at problems that are unique to each of the protocols, concurrent verbalisation discovers 47 problems that are unique to that protocol, whereas retrospective verbalisation only encounters 33 problems that are unique to that protocol (see table 2). It has long been debated in the literature whether unique problems were real or “false” problems, since they had only been encountered by one participant during the evaluation, and how this seems increasingly likely when the number of participants increase. If unique problems are indeed “false” problems, then this experiment could indicate that retrospective verbalisation is better at eliminating these “false” problems. This could be because the protocol is of a recall-nature, where the participant simply recalls fewer of these “false” problems afterwards than what would be verbalised in the situation, due to it not really being a problem after all.

“False” Problems – Do They Exist?

However, retrospective verbalisation finds only slightly more than half of the total number of problems, and the question is if nearly half of the problems found can be considered “false” problems. When looking at the severity, concurrent verbalisation finds more problems in all three categories. If the problems found extra by concurrent verbalisation were “false” problems, it would be fair to assume that they would appear mostly as cosmetic problems. However it is difficult to dismiss problems that are categorised as critical as being false, so eliminating “false” problems can only partly explain why retrospective verbalisation finds only slightly more than half the problems. Another explanation might be that the participant forgets some of the problems in the short time between the evaluation and the retrospective verbalisation. Perhaps problems seem less frustrating when looking back, than when in the middle of it. It is possible that it is easier for the participant to keep the overview when sitting outside the situation looking in.

Problems Detected by Both Protocols

There are 43 problems that are registered by both protocols. As an example one problem was that the participant did not enter username and password before pressing the “login”-button. In another problem the participants did not understand the error message displayed to them. Thirdly, the participants think “Unplanned task” adds an extra task to the visit in progress. These three problems are typical for the 43 problems in common of the two protocols and the initial inspection does not reveal any connection between them that explains why exactly those problems have been revealed by both protocols. The same is the case with the unique problems that also doesn’t seem to have anything in common. Examples of those are: The participant thinks TAB will move the cursor to the next text field. Secondly, a participant is unsure how to end a visit in progress. Thirdly, a participant is unsure what data the “search”-button searches in.

Few or Many – Nothing in Between

It is notable that in concurrent verbalisation it seems like the participants fall in one of two groups. They either experience few or many problems and not the average in between, whereas the number of problems experienced by the participants in retrospective verbalisation is more evened out. Three of the participants using concurrent verbalisation experience only few problems (6-11) while the other five experience many (21-36), but none of the participants experience the average number of problems in between (12-20). This could be due to difficulties in verbalising concurrently with the task-solving, as has been reported as a drawback of the concurrent think-aloud protocol [7]. This can materialise itself either as very little verbalisation due to difficulties doing that simultaneously with the task-solving (few problems experienced) or by extra problems occurring due to lack of concentration caused by the simultaneous verbalisation (many problems experienced). In retrospective verbalisation this is much

more evened out, because the mental workload is lowered by letting the participants concentrate on one thing at a time and the differing number of problems experienced might simply be caused by their varying computer skills and also differing skills in recalling their thought process at the time in details.

The Diverse Participants

Each participant in concurrent verbalisation revealed an average of 20.8 problems, whereas each participant in retrospective verbalisation only discovered an average of 16.0 problems (see table 3). This difference is not particularly big though when considering the large spread in experienced problems between the participants, and this spread is probably to be expected in a group of participants as diverse as the present one. The group contained a wide variety both in job experience and computer experience and as such it would have come as a surprise if the amount of problems experienced were similar between the participants.

Table 3. Average number of problems experienced totally and for each of the two protocols.

	Total	Concurrent Verbalisation	Retrospective Verbalisation
Average problems	18.5	20.8	16.0

Discussion

Many attempts have been made to determine which of the two verbalisation protocols are better, but so far the results are differing between studies. Nielsen et al. [7] discover quite a few weaknesses in concurrent verbalisation, and propose that Mind Tape (a version of retrospective verbalisation) is a more viable option, whereas van den Haak et al. [9] rate the two protocols as being equally good although clearly different. This study indicates that concurrent verbalisation finds more problems than retrospective verbalisation, but it seems that this can be both a good and a bad thing. Good, if it means that the number of "false" problems (unique) is minimized; bad since it is not only "false" problems that aren't discovered. Concurrent verbalisation on the other hand seems to lay a higher mental workload upon the participant, causing them to focus either on the task-solving process and thus tend to forget to verbalise or to focus on the verbalisation thus loosing concentration on the task-solving. However, the reason that retrospective verbalisation finds less problems might be that even in the short time between the actual evaluation and the retrospective verbalisation, things have already started to fade in the memory of the participant and problems are being forgotten. Thus, the conclusion tends to lean towards that of van den Haak et al. [9] that they are equally good, but very different.

As the observant reader might have noticed, the two protocols in the experiment had an uneven number of participants: 8 participants used concurrent verbalisation, while only 7 participants used retrospective verbalisation. This of course influences the results in the subsection *Problems Revealed* of the *Results*-section, but even if the

numbers are corrected to compensate for that (done by taking all possible combinations of 7 participants out of the 8 and then taking of the average of the amount of problems found by these combinations of 7 participants in concurrent verbalisation), concurrent verbalisation still reveals 81.125 problems to retrospective verbalisations 61. This is still a notable difference and does not change the conclusions drawn. The same is the case in the subsection *Unique Problems* where concurrent verbalisation still finds 27.3 of the globally unique problems (compared to the 30) and 41.1 problems that are unique to that protocol (compared to 47) when the numbers are corrected to compensate for the extra participant as described above. Here the differences too are still noteworthy even after the compensation and therefore does not change any of the above written. It of course looks a bit odd to be talking about a fraction of a problem, but it is simply to illustrate the average amount of problems that would have been experienced, if we had only used 7 participants and not 8, regardless which 7 participants we were to choose of the 8. With the corrected numbers, table 2 would then look as can be seen in table 4.

Table 4. Table 2 as it would look with the corrected numbers for concurrent verbalisation.

	All	Concurrent Verbalisation	Retrospective Verbalisation
Problems revealed	105	81.125	61
Unique problems*	44	27.3 (41.1)	14 (33)
Average computer experience	3.4	3.9	3.0

* Note that the number in parentheses refers to problems that are unique to that protocol and not necessarily unique in total.

One purpose of the experiment conducted was to look at the suitability of the protocols for sensitive settings, in this case healthcare in a field evaluation: Surprisingly, and contrary to expected, there was no evidence that the participants using concurrent verbalisation were influenced by the awkwardness or private nature of the information they were verbalising about. This indicates that this is not an issue that affects the test situation or the participant. It is however unclear if this goes for other settings and it would be interesting to explore if, what can be described as sensitive settings, influence the suitability of verbalisation. However, this requires a definition of what makes a sensitive setting, such as surroundings, participants etc., and then identifying application areas where this could pose a problem.

Acknowledgements

The research behind this paper was partly financed by the Danish Research Councils (grant number 2106-04-0022, the USE-project), without which it would not have been possible. I would also like to thank my supervisor for his continuously constructive comments on the paper. Finally, a thank you to the home healthcare workers of Aars kommune in Denmark, who agreed to participate in this experiment, and to the elderly citizens, who so willingly opened their homes to us.

References

1. Dix, A., Finlay, J., Abowd, G. and Beale, R.: Human-Computer Interaction, Prentice Hall (1997)
2. Duncker, K.: On Problem-solving, in Dashiell, John F.: Psychological Monographs, The American Psychological Association, Inc., vol. 58, pp. 1-114 (1945)
3. Ericsson, K. A. and Simon, H. A.: Protocol Analysis: Verbal Reports as Data. Cambridge, MA: MIT Press (1993)
4. Hackos, J. T. and Redish, J. C.: User and Task Analysis for Interface Design, Wiley (1998)
5. Molich, R.: User Friendly Systems (in Danish), Teknisk Forlag (1994)
6. Nielsen, J.: Estimating the Number of Subjects Needed for a Thinking Aloud Test, International Journal of Human-Computer Studies, vol. 41, no. 3, pp. 385-397 (1994)
7. Nielsen, J., Clemmensen, T. and Yssing, C.: Getting access to what goes on in people's heads? – Reflections on the think-aloud technique. In: Proceedings of NordiCHI. ACM, (2002)
8. Preece, J.: Human-Computer Interaction, Addison-Wesley (1994)
9. van den Haak, M., de Jong, M. D. T. and Schellens, P. J.: Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. In: Behaviour and Information Technology, Vol. 22, No. 5, pp 339-351 Taylor & Francis Ltd. (September-October 2003)

A Classification of Research Methods and Purposes in Child-Computer Interaction

Janne J. Jensen, Mikael B. Skov
Department of Computer Science
Aalborg University
Fredrik Bajers Vej 7
DK-9220 Aalborg East, Denmark
{jjj, dubois}@cs.aau.dk

ABSTRACT

Research methods have been objects of discussions for decades and defining research methods is still a substantial challenge. However, it is important to understand how research methods have been adapted in different disciplines as it potentially informs us on future directions and influences on the discipline. Inspired by previous studies from other disciplines, we conduct a classification of research methods in paper publications within child-computer interaction (CCI). 132 papers on CCI are classified on a two-dimensional matrix on research method and purpose. Our results show a strong focus on engineering of products as applied research and on evaluation of developed products in the field or in the lab. Also, we find that much research is conducted in natural setting environments with strong focus on field studies. Finally, gender issues are important in many research studies with children while age issues play less significant roles.

KEYWORDS

Research methods, children's technologies, HCI

1. INTRODUCTION

During the past decade, there has been a significant increase in the published work relating to children and interaction design and the annually held IDC conferences strongly support this growing development. As a consequence, we are currently seeing that child-computer interaction (CCI) is becoming a vibrant sub-field of human-computer interaction (HCI). While HCI has been growing in importance during the last decades and has matured as a discipline, CCI is still rather immature and finding its way. In several ways, CCI is truly multi-disciplinary and integrates elements from education and educational technology, and connects to art, design, storytelling, and literature. This disparity in methods of enquiry makes it difficult for researchers to gain an overview of research, compare across studies and gain a clear view on cumulative progress in this field. Different research methods have been adapted in research projects involving children. This is no different than other disciplines, but it is important to understand how research methods have been adapted in different disciplines as it potentially informs us on future directions and influences on the discipline (Kjeldskov and Graham, 2003).

Inspired by studies within information systems and related disciplines, we wish to evoke the discussion of research methods adapted in CCI during the last decade. Research methodology has been examined in information systems for years; see (Galliers, 1990; Myers, 1997; Wynekoop and Congor, 1990). A number of frameworks have been proposed to facilitate the discussion of research methods in information systems. For the study in this paper, we find the classification scheme found in Wynekoop and Congor (1990), useful as it provides a simple but powerful analysis of a research discipline. We wish to provide a snapshot of current and previous research conducted within our discipline to high-

Table 1. Summary of research methods on strengths, weaknesses, and use (adapted from Wynekoop and Congor, 1990 and Kjeldskov and Graham, 2003).

	Method	Strengths	Weaknesses	Use
Natural Setting	Case studies	Natural settings Rich data	Time demanding Limited generalizability	Descriptions, explanations, developing hypothesis
	Fields studies	Natural Settings Replicable	Difficult data collection Unknown sample bias	Studying current practice Evaluating new practices
	Action research	First hand experience Applying theory to practice	Ethics, bias, time Unknown generalizability	Generate hypothesis/theory Testing theories/hypothesis
Artificial Setting	Laboratory experiments	Control of variables Replicable	Limited realism Unknown generalizability	Controlled experiments Theory/product testing
Environment Independent Setting	Survey research	Easy, low cost Can reduce sample bias	Context insensitive No variable manipulation	Collecting descriptive data from large samples
	Applied research	The goal is a product which may be evaluated	May need further design to make product general	Product development, testing hypothesis/concepts
	Basic research	No restrictions on solutions Solve new problems	Costly, time demanding May produce no solution	Theory building
	Normative writings	Insight into firsthand experience	Opinions may influence outcome	Descriptions of practice, building frameworks

light how the research has been carried out. Thus, our aim is also to provide a mechanism that can be used to further develop a community of researchers within CCI, which is likely as important to a young discipline as ours. Section two outlines and describes the classification matrix explaining different research methods and purposes. Section three classifies research methods in papers on CCI (the papers are listed in the appendix). Section four discusses the results of the study and compares the results to studies of other disciplines.

2. CLASSIFICATION OF RESEARCH METHODS

Research methods have been objects of discussions for decades and defining research methods is still a quite substantial challenge (Kjeldskov and Graham, 2003). Since the aim of this paper is to classify existing research papers according to applied research methods in the design of children's technologies, it is not our intention to define research methods or propose new research methods. As a result, we have chosen a definition found in Wynekoop and Congor (1990) for classification of research methods in computer aided software engineering (CASE) and later adapted by Kjeldskov and Graham (2003) for mobile human-computer interaction research methods.

This classification of research methods proposes a matrix of two dimensions namely research methods and research purposes. In the following, we will provide a description of the research methods and purposes extracted from the discussions in (Kjeldskov and Graham, 2003; Wynekoop and Congor, 1990), supplemented by definitions and discussions of research methods in information systems (Rapoport, 1970; Stone, 1981; Yin, 1994) (for more detailed descriptions please refer to (Wynekoop and Congor, 1990, pp. 132-141) or (Kjeldskov and Graham, 2003; pp. 318-324).

2.1. RESEARCH METHODS

The eight research methods include case studies, field studies, action research, lab experiments, survey research, applied research, basic research, and normative writings. The first three are natural setting research methods conducted in real organizational settings, the fourth is an artificial setting research

method conducted in a laboratory, while the latter four are environment independent setting research methods as they assume no influence by the context of the conduction.

The natural setting research methods are conducted in real organizational settings and include case studies, field studies, and action research. *Case studies* are intensive evaluations of small samples of entities e.g. groups, organizations, individuals, systems, or tools (Yin, 1994). Usually researchers will collect both quantitative and qualitative data through multiple means including interviews, observation, questionnaires etc. Often none or few experimental or statistical controls are enforced (Galliers, 1990). Case studies have been found particularly useful for explaining or describing phenomena and for developing hypotheses, but they can be rather time consuming and generalization of findings is sometimes limited. *Field studies* are research activities taking place in the real world (opposed laboratory environments). Field studies can integrate both quantitative and qualitative approaches ranging from ethnographic studies to field experiments. Ethnographic field studies typically bring the researcher in the field spending considerable time observing the environment, whereas field experiments are characterized by manipulation of independent variables to observe changes in a natural setting (Galliers, 1990). One advantage of field studies is that they often yield results over a relative short period of time, but researchers face a risk of experimental manipulation in field experiments. *Action research* reflects research where the researcher conducts the research activities while participating in the intervention and simultaneously evaluating the results (Myers, 1997). Action research aims at both contributing to the practical concerns of people in problematic situations and to the goal of social science in a joint collaboration (Rapoport, 1970). Action research has some advantages, e.g. the researcher gains a first-hand understanding of the situation and the researcher is not viewed as interfering with the process. The drawbacks are that action research is rather time consuming and ethical challenges emerge as researcher gains understanding of phenomena while at the same time concealing them.

The artificial setting research method is conducted in a laboratory and includes laboratory experiments. *Lab experiments*, opposed to field experiments, take place in a controlled environment with the experimenter in control of assignments of subjects, treatment variables, and manipulation of variables (Stone, 1981). Major advantages of lab experiments are more precise measurements of the phenomena studied and enhanced possibilities to replicate. Disadvantages include that generalization is limited to the sample population and the assumption that real-world interference is not important.

The environment independent setting research methods are assumed to have no influence by the context of the conduction and include survey research, applied research, basic research, and normative writing. 1) *Survey research* applies information from a known population gathered through e.g. interviews or questionnaires. The data is collected directly from the respondents and normally assumes unaffected by the context. The advantages of surveys are that very large samples can be collected in a relative short period of time and generalization can be achieved for a broader population. Disadvantages include the assumption of snapshot of phenomena which often requires triangulation of different approaches. 2) *Applied research* informs research where intuition, experience, deduction, and induction are used to analyze a specific research problem (Wynekoop and Congor, 1990). Typically, the approach taken in applied research to solution finding is trail and error based on the capabilities of the researcher. One advantage of this approach is the goal-directness and the usefulness of an end-product being developed. Drawbacks include that the initial solution may not be elegant, easily adapted, or context-independent. 3) *Basic research* is about developing new theories or performing research in a field where the problem is known, but the methods and solutions are not known. The approach is, like applied research, trail and error based on the capabilities of the researcher. Basic research holds the advantage that no preconceptions exist and there is often no time pressure. Disadvantages are that the research is slow and may not yield any useful solutions. 4) *Normative writing* is a final category of research methods included by Wynekoop and Congor which they refer to as "non-research" writings about phenomena of interests. They suggest that normative writings include concept development writings, presentation of "truth", and application descriptions (Wynekoop and Congor, 1990). Concept developments indicate direction for future research whereas presentations of "truth" present ideas of con-

Table 2: Selected outlets, numbers of candidate papers, and numbers of selected papers

Outlets: Journals/Proceedings (1996-2005)	Candidate papers N	Candidate papers %	Selected papers N	Selected papers %
ACM Transactions on Human-Computer Interaction (TOCHI)	135	4	3	2
International Journal of Human-Computer Studies (IJHCS)	684	21	3	2
International Journal of Human-Computer Interaction (IJHCI)	258	8	0	0
Behaviour and Information Technology (BIT)	356	11	3	2
Interacting with Computers (IwC)	326	10	9	7
Personal and Ubiquitous Computing (PUC)[1]	214	6	10	8
Conference on Human Factors in Computing Systems (CHI)	732	22	40	30
Conference on Interaction Design and Children (IDC)[2]	54	2	55	42
International Conference on Human-Computer Interaction (Interact)[3]	381	12	8	6
Symposium on Designing Interactive Systems (DIS)[4]	155	5	1	1
Total	3295	100	132	100

[1] *Personal and Ubiquitous Computing has been published since 2000*
[2] *The IDC conference has been held annually since 2002*
[3] *The Interact conference has been held bi-annually (1997, 1999, 2001, 2003, 2005)*
[4] *The DIS conference has been held four times in the period (1997, 2000, 2002, 2004)*

cepts that seem intuitively correct. Application descriptions are narratives written by practitioners outlining subjective views on situations or phenomena. The main advantage of normative writings is that they require very little effort to produce.

2.2. RESEARCH PURPOSES

Wynekoop and Congor (1990) propose a second dimension in their matrix namely research purpose. In our review of research methods in children's technology design, we will adapt the same dimension. The categories and definitions for the five research purposes are summarized below. 1) *Understand* is the focus on grasping the meaning of the entities being studied, e.g. frameworks that attempts to categorize for better understanding. 2) *Engineer* is the focus of research where the aim is to develop new systems or parts of systems. 3) *Re-engineer* is the re-development of an existing system or part of a system usually based on an evaluation. 4) *Evaluate* is the assessing or validation of a product or a system, either to compare a single product or to compare more products. 5) *Describe* is the research that defines or describes features of an ideal system or situation.

3. CLASSIFICATION OF RESEARCH METHODS

This section will present our classification of research methods in selected research papers on design of children's technologies. This will be done accordingly to the definitions of the matrix by Wynekoop and Congor (1990). A total of 132 papers were selected from the following top-level journals and conference proceedings for the period 1996-2005 (see table 2). While other journals and conferences exist, we find that the pool of included research papers provides a solid base for the classification given the number of papers and the high-quality reviewing process of the journals and conferences. The 132 papers were selected for the review based on a thorough examination of all full research publications in the above journals and proceedings.

During the period (1996-2005), a total of 3295 candidate papers were published in the selected outlets (se table 2). We read abstracts (and occasionally introductions, methods etc.) of all 3295 candidate papers and a paper was selected for the classification if it dealt with issues or aspects of children's technology design. The 132 selected papers were printed out, numbered, read through. Like Kjeldskov and Graham (2003), we aimed to ensure consistency by scanning all papers a second time over a few days

Table 3. Classification of research methods in design of children's technologies. The numbers refer to the items listed in the appendix of the reviewed research papers.

	Case studies	Field studies	Action research	Lab experiment	Survey research	Applied research	Basic research	Normative writings
Understand	5, 19, 20, 43, 46, 47, 65, 72, 79	27, 29, 30, 48, 49, 55, 60, 61, 71, 74, 81, 93, 102, 111, 113, 129		4, 8, 23, 26, 55, 107, 113, 126	63, 77, 114			34
Engineer		22, 32, 44, 58, 84, 90, 91, 118, 121		2, 31, 32, 33, 85		5, 6, 12, 13, 17, 18, 19, 24, 25, 27, 37, 38, 42, 45, 47, 48, 51, 52, 60, 64, 66, 67, 70, 73, 78, 80, 86, 87, 88, 89, 94, 95, 96, 97, 99, 100, 102, 104, 106, 108, 109, 110, 112, 113, 114, 115, 116, 117, 119, 120, 121, 123, 124, 130, 131, 132		
Re-engineer		118		53, 54, 68		11, 24, 51, 52, 78, 94		
Evaluate		9, 10, 11, 12, 15, 16, 17, 18, 19, 22, 25, 28, 29, 30, 31, 37, 39, 42, 47, 48, 52, 59, 62, 66, 67, 73, 74, 76, 78, 82, 83, 89, 90, 91, 92, 95, 99, 100, 106, 109, 110, 112, 114, 119, 120, 121, 127, 130, 131, 132		3, 4, 5, 6, 7, 14, 21, 23, 24, 38, 44, 50, 51, 52, 53, 54, 56, 57, 64, 69, 75, 80, 85, 88, 89, 94, 96, 97, 98, 101, 104, 105, 107, 112, 113, 117, 122, 123, 124, 128				
Describe							8	1, 35, 36, 40, 41, 62, 82, 83, 84, 86, 103, 106, 112, 120, 125

and to ensure validity by having both authors reading and classifying all 132 papers individually and then afterwards negotiate the classifications in collaborative effort. The classification of the papers can be found in table 3. As with the survey by Kjeldskov and Graham (2003), some of the reviewed papers clearly employed more than one research method and had multiple purposes. As a consequence, these papers were given multiple classifications and appear more than once in the table. This implies that aggregate percentages sometimes amount to more than 100%.

Table 3 shows that 58% of the selected papers fall into the field study category (76 of 132 papers). The second and third most used categories are applied research (47%) and lab experiments (42%). We found 16 entries for normative writings, nine for case studies, three surveys, and one for basic research, and zero for action research. Our study indicates no clear bias towards either natural setting environ-

ments, artificial setting environments, or independent setting environments, but there is a somewhat strong focus on natural setting environments.

Considering the research purpose, we find that 68% of the papers did some sort of evaluation (90 of 132 papers), of which 56% are conducted in field evaluations (50 of 90) and 44% are conducted in laboratory experiments (40 of 90). The second most preferred purpose is engineering with 52% of the papers (69 of 132 papers) of which 81% would employ applied research as research method (56 of 69). Also, 28% papers fall into the category of understanding mostly based on case studies, field studies, or lab experiments. Thus, there seems to be a clear bias towards evaluating products (often with children at different ages, but also different kinds of adults, e.g. teachers) and on developing (engineering) prototypes and products for children.

Of the 62 papers on applied research, 90% would do so for engineering purposes while 10% would re-engineer. Considering the 56 papers employing applied research for engineering purposes, we found that 52% of these papers followed up on their design with a field evaluation and another 36% followed up with a laboratory evaluation (three papers conducted both a field and a lab evaluation). Hence, 10 of the 56 papers (18%) did not report from any evaluation of the engineered solution. Furthermore and quite remarkably, only seven of the 56 papers also *report* from activities on understanding, thus 88% did not *report* any findings related the research purpose understand.

Many of the papers involve research in natural setting environments with 64% (85 of 132 papers) and most of this research takes place in field studies 89% (76 of 85). Furthermore, of the 37 papers aiming to understand, most would conduct research in a natural setting environment (68%). This would usually be done be observing children in their natural habitat, e.g. schools. On the other hand, 22% of the understanding papers would employ a lab-based setup (8 of 37 papers). Finally, of the 16 papers in the description category, 15 would fall into the normative writing category proposing ideas and suggestions of e.g. methods for developing with children.

4. DISCUSSION

Our study reveals that much research on child-computer interaction (CCI) focus on evaluating or engineering purposes and many papers present some design solution typically followed by a controlled, systematic evaluation with the purpose of assessing the success of the engineered solution. On the other hand, there is no clear bias towards any preferred environment for research conduction on children's technology design, but natural setting environments are commonly used. Such research is typically conducted in schools primarily for evaluating educational products. Examining the results of our survey further, we identify a number of issues that seem to characterize the CCI field. In the following, we refer to some of the research papers in the appendix through numbers in brackets e.g. [23].

Our classification reveals that our discipline has a rather strong focus on natural setting environments. This is pursued primarily through different kinds of field studies and secondarily through case studies. The strong focus on natural settings and field studies is in deep contrast to the survey study on mobile technologies. Kjeldskov and Graham (2003) found that for research on mobile technologies very few studies moved into a real world context for any research purpose. One of the identified problems was the immediate lack of control in a real world setting, e.g. when evaluating a product in the field it could be difficult to judge influence of contextual factors when assessing the mobile system. However, this lack of control does not seem to influence the evaluation setting for many studies on children's technology design as many would evaluate their design solution through a field study evaluation. Rather than viewing the dynamics of the real world context as problematic, more research studies on children's technology design regard this influence as useful and necessary for understanding the usefulness and usability of the produced solution. Furthermore, the strong focus on field studies may also come from the fact that when evaluating children's technologies the most obvious way to recruit subjects is to place the evaluation in a school environment and several evaluations take place in schools, e.g. [78, 99, 120, 121, 131]. Other studies exploit the field as natural component of evaluating

context-aware or learning technologies that are closely related to the context, e.g. [90, 114]. On the other hand, we identified no studies employing action research as research method. This is comparable with the mobile technology survey (Kjeldskov and Graham, 2003). Kjeldskov and Graham state that the lack of action research is due to a rather limited established body of theoretical knowledge and an unwillingness to implement these technologies in real life mainly due to high costs. This could also be the case for our discipline, but the included research papers provide no clear indication on the lack of action research.

The classification demonstrated only limited focus on understanding as research purpose. This is somewhat surprising as children are generally acknowledged to have different needs and capabilities in relation to software technologies compared adults. As a consequence, a major research challenge in CCI remains to better understand children as users and consumers. However, our study found a rather limited focus on reporting issues of understanding. From the papers focusing on understanding, we found a tendency towards that researchers would to utilize case studies when they want to understand more general level issues of children's use and perception of software technologies, e.g. impact of using natural language programming languages [20], social impact of technologies [43], and unique needs of pre-school children in learning environments [46]. Field-based studies or lab-based experiments are often conducted when researchers are trying to investigate specific aspects of children's interaction with software technologies. Field studies are then often applied when the context of use plays an important role, e.g. involvement and learning in a museum [48], impact of a distributed, cooperative system on the educational practices in a school [27], or relative benefits of two data gathering techniques [102]. On the other hand, lab experiments are often conducted to understand relative benefits of exiting or emerging methods, e.g. think-aloud and constructive interaction [3, 4]. Despite these studies, understanding as research purpose is very little represented compared engineering and evaluation.

Given the strong focus on applied research for engineering purposes, it seems quite surprisingly that very few papers also *report* from research purposes of understanding. This lack of focus was also identified for mobile technologies and Kjeldskov and Graham (2003) concluded that the question of usefulness and what is perceived to be problematic from a user perspective is poorly represented in mobile technology research. The limited focus on understanding prohibits us from a deeper understanding of the needs and requirements of children in relation to use of new technologies. Such information could potentially inform us on new and innovative products for children. On the other hand, only seven papers *report* from understanding as research purpose when also engineering a product as applied research, e.g. [5, 19, 27, 47, 60, 113, 114]. This does not necessarily imply that the other studies on applied engineering did not conduct activities related understanding, but these papers did not reflect such activities. Furthermore, the limited focus on case studies and survey research prohibits our discipline from research results that could collect large amounts of data from, for example, children's actual use of current technologies and more general preferences of contemporary and future technologies.

Very few papers explicitly reflect upon issues related the age of the involved children. One can argue that children are primarily defined by their age; age has significant influence on children's capabilities and skills. Our study focused on research involving children up to 18 years and the selected studies involved children at all ages ranging from pre-school children, such as children aged 2-5 in the design of ActiMates [123], through middle-aged children, such as Audiodoom for visually impaired children aged 7-11 [80], to older children, such as the Progress Portfolio tool for children aged 14-18 [78]. However, most studies involve middle-age children (8-12 years old) whereas much fewer studies involve pre-school children and teenagers. Furthermore, some research studies focused on or involved children using a broad age range, e.g. the programming tool for girls aged 6-13 [47], while other studies focused on a much more narrow age range, e.g. HandLeR a mobile educational device involving children aged 10-11 [114]. Somewhat surprisingly, only few studies explicitly reflect upon the children's age and their involvement in interaction design. Wyeth and Wyeth state that pre-school children can be engaged with computers given the right interfaces [130], and Raffle et al. report on age range findings for Topobo, a 3D constructive assembly system [99]. They found that kindergarten children

(5-6 years old) would engage differently with the developed system compared second graders (7-8 years old) and more focus on single aspects of the system. Finally, Bekker et al state that usability evaluators should phrase tasks carefully according to the age of the children [9, 10]. Ling conducts a larger survey among children and teenagers in Norway and reports on differences and similarities between children use of mobile phones for different ages [77].

We found that much CCI research focus on gender issues; opposed other research discipline classifications (Kjeldskov and Graham, 2003; Wynekoop and Congor, 1990). Several of the selected research studies involved only girls in the process, e.g. [32, 33, 37, 47, 60, 88, 104], while surprisingly we found no studies involving boys only. The primary motivation for involving only girls in the studies seems to stem from the fact that girls are poorly represented in major computer science and information technology educations in both the States and in Western Europe. As an example, Gweon et al. deliberately attempted to expose girls to programming through creative tools [47] while Isomursu et al. involve girls only in a design process as they state that young girls are often neglected in the design of technical devices [21]. Other studies report on the impact of gender in the design process, e.g. Als et al. found that acquainted dyads of boys collaborated better than acquainted dyads of girls when usability evaluating using constructive interaction [4] while Fails et al found that pre-school girls verbalized more than boys during interaction with physical and desktop environments [38]. Stringer et al. found that boys showed higher enthusiasm for a given technology than girls which gave the boys a way into the activity on more equal terms [121]. Ling found that girls and boys have different mobile phone traditions where e.g. girls would more often borrow mobile phones when living at home than boys [77]. Hourcade et al. coincidently found a number of gender issues in pre-school children. They found that boys performed better than girls in some pointing task assignments [55]. Also, Benford et al. discovered that children would often team up in gender-wise pairs if given the opportunity [12]. The TurboTurtle project adapted mixed gender pairs to explore male domination during collaboration and found surprisingly rather extreme behaviours of the children with respect to collaboration in mixed gender pairs [24].

5. CONCLUSION

We are currently seeing that child-computer interaction (CCI) is becoming a vibrant sub-field of human-computer interaction (HCI) and while HCI has been growing in importance during the last decades and has matured as a discipline, CCI is still rather immature and finding its way. As a modest attempt to mature our discipline, we conducted a classification of 132 research papers on child-computer interaction to highlight research methods and purposes. Our classification showed that CCI has a rather strong focus on natural setting environments primarily pursued through field studies and secondarily through case studies. Also, we found a rather limited focus on reporting issues of understanding within CCI as most papers would report on engineering or evaluation purposes. Further, age seems to play a double-sided role in CCI as it is generally acknowledged as an important aspect, but rather few studies explicitly investigate age aspects or reflect upon this aspect. Finally, gender plays an important role in CCI opposed other research disciplines.

The classification provides a number of opportunities for future research within our discipline. First, the tight integration of children and designers/researchers could be further explored in action research projects. Secondly, different kinds of research on, for example, case studies and survey could inform us on different issues and from different perspectives on children's use of technologies.

Our classification is limited in a number of ways. First, the classification matrix was designed for and built upon research in the field of information systems. Thus, the applicability of the matrix for CCI research may be limited. As it can be seen from the classification table, several of the cells are empty and the combination of some methods and purposes may be infeasible. Secondly, classifying research papers according to methods and purposes was difficult as many papers would fall into more categories, and as several papers lacked information on research methods and purposes. Furthermore,

the descriptions of adapted methods were often ambiguous resulting in several possible interpretations. As a result, we had to read and review the papers in several iterations; also we renegotiated definitions of methods, e.g. action research.

ACKNOWLEDGEMENTS

The work behind this paper received financial support from the Danish Research Agency (grant no. 2106-04-0022). We would like to thank Alissa Antle for constructive comments on an earlier version of the paper and more anonymous reviewers for comments on earlier drafts of the paper.

REFERENCES

Galliers, R. D. (1990) Choosing Appropriate Information Systems Research Approaches: A Revised Taxonomy. In *Proceedings of the IFIP TC8 WG8.2 Working Conference on the Information Systems Research Arena of the 90's*, Copenhagen, Denmark

Kjeldskov, J. and Graham, C. (2003) A Review of MobileHCI Research Methods. In *Proceedings of the 5th International Conference on Mobile Human-Computer Interaction*, LNCS, pp. 317 - 335

Myers, M. D. (1997) Qualitative Research in Information Systems. *MIS Quarterly*, Vol. 21(2), pp. 241 - 242

Rapoport, R. N. (1970) Three Dilemmas in Action Research. *Human Relations*, Vol. 23(4), pp. 499 - 513

Stone, E. (1981) *Research Methods in Organization Behavior*. Scott-Foresman, Chicago

Wynekoop, J. L. and Congor, A. A. (1990) A Review of Computer Aided Software Engineering Research Methods. In *Proceedings of the IFIP TC8 WG8.2 Working Conference on the Information Systems Research Arena of the 90's*, Copenhagen, Denmark

Yin, R. K. (1994) *Case Study Research, Design and Methods*, Newbury Park, 2nd edition, Sage Publications

APPENDIX: REVIEWED RESEARCH PAPERS 1996-2005

1. Ackerman, E. K. (2005) Playthings that do Things: A Young Kid's 'Incredibles'! In Proceedings of the Conference on Interaction Design and Children IDC'05
2. Alborzi, H., Druin, A., Montemayor, J., Platner, M., Porteous, J., Sherman, L., Boltman, A., Taxén, G., Best, J., Hammer, J., Kruskal, A., Lal, A., Schwenn, T. P., Sumida, L., Wagner, R. and Hendler, J. (2000) Designing StoryRooms: Interactive Storytelling Spaces for Children. In Proceedings of Symposium on Designing Interactive Systems, pp. 95-104
3. Als, B. S., Jensen, J. J. and Skov, M. B. (2005) Comparison of Think-Aloud and Constructive Interaction in Usability Testing with Children. In Proceedings of the Conference on Interaction Design and Children IDC'05
4. Als, B. S., Jensen, J. J. and Skov, M. B. (2005) Exploring Verbalization and Collaboration of Constructive Interaction with Children. In Proceedings of the IFIP TC13 Interact '05, IOS Press, pp. 443 - 456
5. Antle, A. (2003) Case Study: The Design of CBC4Kids' Storybuilder. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 59 - 68
6. Antle, A: (2004) Supporting children's emotional expression and exploration in online environments. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 97 - 104
7. Baauw, E., Bekker, M. M. and Barendregt, W. (2005) A Structured Expert Evaluation Method for the Evaluation of Children's Computer Games. In Proceedings of the IFIP TC13 Interact '05, IOS Press, pp. 457 - 469
8. Barendregt, W., Bekker, M. M. and Speerstra, M. (2003) Empirical Evaluation of Usability and Fun in Computer Games for Children. In Proceedings of the IFIP TC13 Interact '03, IOS Press, pp. 705 - 708

9. Bekker, M, Beusmans, J., Keyson, D. and Lloyd, P. (2002) KidReporter: A Method for Engaging Children in Making a Newspaper to Gather User Requirements. In Proceedings of the Conference on Interaction Design and Children IDC'02
10. Bekker, M, Beusmans, J., Keyson, D. and Lloyd, P. (2003) KidReporter: A User Requirements Gathering Technique for Designing with Children. Interacting with Computers, Vol. 15, pp. 187-202
11. Benford, S., Bederson, B. B., Åkesson, K-P, Bayon, V., Druin, A., Hansson, P., Hourcade, J. P., Ingram, R., Neale, H., O'Malley, C., Simsarian, K. T., Stanton, D., Sundblad, Y. and Taxén, G. (2000) Designing Storytelling Technologies to Encourage Collaboration between Young Children. In Proceedings of the Conference on Human Factors in Computing Systems CHI'00, ACM, pp. 556 - 563
12. Benford, S., Rowland, D., Flintham, M., Drozd, A., Hull, R., Reid, J., Morrison, J. and Facer, K. (2005) Life on the Edge: Supporting Collaboration in Location-Based Experiences. In Proceedings of the Conference on Human Factors in Computing Systems CHI'05, ACM, pp. 721 - 730
13. Berglin, L. (2005) Spookies: Combining Smart materials and Information technology in an Interactive Toy. In Proceedings of the Conference on Interaction Design and Children IDC'05
14. Bernard, M. L., Chaparro, B. S., Mills, M. M. and Halcomb, C. G. (2002) Examining Children's Reading Performance and Preference for Different Computer-Displayed Text. Behaviour and Information Technology, Vol. 21 (2), pp. 87-96
15. Bers, M. U., Ackermann, E., Cassell, J., Donegan, B., Gonzalez-Heydrich, J., DeMaso, D. R., Strohecker, C., Lualdi, S., Bromley, D. and Karlin, J. (1998) Interactive Storytelling Environments: Coping with Cardiac Illness at Boston's Children's Hospital. In Proceedings of the Conference on Human Factors in Computing Systems CHI'98, ACM, pp. 603-610
16. Bers, M. U., Gonzalez-Heydrich, J. and DeMaso, D. R. (2001) Identity Construction Environments: Supporting a Virtual Therapeutic Community of Pediatric Patients Undergoing Dialysis. In Proceedings of the Conference on Human Factors in Computing Systems CHI'01, ACM, pp. 380 - 387
17. Borovoy, R., Silverman, B., Gorton, T., Klann, J., Notowidigdo, M., Knep, B. and Resnick, M. (2001) Folk Computing: Revisiting Oral Tradition as a Scaffold for Co-Present Communities. In Proceedings of the Conference on Human Factors in Computing Systems CHI'01, ACM, pp. 466 - 473
18. Bouvin, N. O., Brodersen, C., Hansen, F. A., Iversen, O. S. and Nørregaard, P. (2005) Tools of Contextualization: Extending the Classroom to the Field. In Proceedings of the Conference on Interaction Design and Children IDC'05
19. Brederode, B., Markopoulos, P., Gielen, M., Vermeeren, A. and de Ridder, H. (2005) pOwerball: The design of a novel mixed-reality game for children with mixed abilities. In Proceedings of the Conference on Interaction Design and Children IDC'05
20. Bruckman, A. and Edwards, E. (1999) Should We Leverage Natural-language Knowledge? An Analysis of User Errors in a Natural-language-style Programming Language. In Proceedings of the Conference on Human Factors in Computing Systems CHI'99, ACM, pp. 207-214
21. Cassell, J. and Ryokai, K. (2001) Making Space for Voice: Technologies to Support Children's Fantasy and Storytelling. Personal and Ubiquitous Computing, Vol. 5, pp. 169-190
22. Chen, C-H., Wu, F-G., Rau, P-L. P. and Hung, Y-H. (2004) Preferences of young children regarding interface layouts in child community web sites. Interacting with Computers, Elsevier, Volume 16(2), pp. 311 - 330
23. Chiasson, S. and Gutwin, C. (2005) Testing the Media Equation with Children. In Proceedings of the Conference on Human Factors in Computing Systems CHI'05, ACM, pp. 829 - 838
24. Cockburn, A. and Greenberg, S. (1998) The Design and Evolution of TurboTurtle, a Collaborative Microworld for Exploring Newtonian Physics. International Journal of Human-Computer Studies, Vol. 48, pp. 777-801
25. Danesh, A., Inkpen, K., Lau, F., Shu, K. and Booth, K. (2001) GeneyTM: Designing a Collaborative Activity for the palmTM Handheld Computer. In Proceedings of the Conference on Human Factors in Computing Systems CHI'01, ACM, pp. 388 - 395
26. Decortis, F. and Rizzo, A. (2002) New Active Tools for Supporting Narrative Structures. Personal and Ubiquitous Computing, Vol. 6, pp. 416-429
27. Decortis, F., Marti, P., Moderini, C., Rizzo, A. and Rutgers, J. (2002) Disappearing Computer, Emerging Creativity: An Educational Environment for Cooperative Story Building. In Proceedings of the Conference on Interaction Design and Children IDC'02

28. Decortis, F., Rizzo, A. and Saudelli, B. (2003) Mediating Effects of Active and Distributed Instruments on Narrative Activities. Interacting with Computers, Vol. 15, pp. 801 - 830
29. Dindler, C., Eriksson, E., Iversen, O. S., Lykke-Olesen, A. and Ludvigsen, M. (2005) Mission from Mars – A Method for Exploring User Requirements for Children in a Narrative Space. In Proceedings of the Conference on Interaction Design and Children IDC'05
30. Donker, A. and Reitsma, P. (2004) Usability testing with young children. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 43 - 48
31. Druin, A., Stewart, J., Proft, D., Bederson, B. and Hollan, J. (1997) KidPad: A Design Collaboration between Children, Technologists, and Educators. In Proceedings of the Conference on Human Factors in Computing Systems CHI'97, ACM, pp. 463 – 470
32. Druin, A. (1999) Cooperative Inquiry: Developing New Technologies for Children with Children. In Proceedings of the Conference on Human Factors in Computing Systems CHI'99, ACM, pp. 592 - 599
33. Druin, A., Montemayor, J., Hendler, J., McAlister, B., Boltman, A., Fiterman, E., Plaisant, A., Kruskal, A., Olsen, H., Revett, I., Schwenn, T. P., Sumida, L. and Wagner, R. (1999) Designing PETS: A Personal Electronic Teller of Stories. In Proceedings of the Conference on Human Factors in Computing Systems CHI'99, ACM, pp. 326 - 329
34. Druin, A. and Inkpen, K. (2001) When are Personal Technologies for Children? Personal and Ubiquitous Computing, Vol. 5, pp. 191-194
35. Eisenberg, M., Eisenberg, A., Hendris, S., Blauvelt, G., Butter, D., Garcia, J., Lewis, R. and Nielsen, T. (2003) As We May Print: New Directions in Output Devices and Computational Crafts for Children. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 31 - 39
36. Eisenberg, M. (2004) Tangible ideas for children: materials sciences as the future of educational technology. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 19 - 26
37. Ellis, J. B. and Bruckman, A. S. (2001) Designing Palaver Tree Online: Supporting Social Roles in a Community of Oral History. In Proceedings of the Conference on Human Factors in Computing Systems CHI'01, ACM, pp. 474 – 481
38. Fails, J. A., Druin, A., Guha, M. L., Chipman, G., Simms, S. and Churaman, W. (2005) Child's Play: A Comparison of Desktop and Physical Interactive Environments. In Proceedings of the Conference on Interaction Design and Children IDC'05
39. Fels, D. I., Waalen, J. K., Zhai, S. and Weiss, P. T. (2001) Telepresence Under Exceptional Circumstances: Enriching the Connection to School for Sick Children. In Proceedings of the IFIP TC13 Interact '01, IOS Press, pp. 617 - 624
40. Fisch, S. M. (2004) What's so "new" about "new media?": comparing effective features of children's educational software, television, and magazines. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 105 – 111
41. Fisch, S. M. (2005) Making Educational Computer Games "Educational". In Proceedings of the Conference on Interaction Design and Children IDC'05
42. Frei, P., Su, V., Mikhak, B. and Ishii, H. (2000) curlybot: Designing a New Class of Computational Toys. In Proceedings of the Conference on Human Factors in Computing Systems CHI'00, ACM, pp. 129 - 136
43. Frohlich, D. M., Dray, S. and Silverman, A. (2001) Breaking Up is Hard to do: Family Perspectives on the Future of the Home PC. International Journal of Human-Computer Studies, Vol. 54, pp. 701-724
44. Gibson, L., Newall, F. and Gregor, P. (2003) Developing a Web Authoring Tool that Promotes Accessibility in Children's Design. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 23 - 24
45. Gorbet, M. G., Orth, M. and Ishii, H. (1998) Triangles: Tangible Interface for Manipulation and Exploration of Digital Information Technology. In Proceedings of the Conference on Human Factors in Computing Systems CHI'98, ACM, pp. 49 - 56
46. Guha, M. L., Druin, A., Chipman, G., Fails, J. A., Simms, S. and Farber, A. (2004) Mixing ideas: a new technique for working with young children as design partners. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 35 - 42
47. Gweon, G., Ngai, J. and Rangos, J. (2005) Exposing Middle School Girls to Programming via Creative Tools. In Proceedings of the IFIP TC13 Interact '05, IOS Press, pp. 431 – 442
48. Hall, T. and Bannon, L. (2005) Designing Ubiquitous Computing to Enhance Children's Interaction in Museums. In Proceedings of the Conference on Interaction Design and Children IDC'05

49. Hammann, E. and Hennessey, J. M. (2002) How to Attract Early Teens to Your Mobile Network. In Proceedings of the Conference on Interaction Design and Children IDC'02
50. Hanna, L., Neapolitan, D. and Risden, K. (2004) Evaluating computer game concepts with children. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 49 - 56
51. Henderson, V., Lee, S., Brashear, H., Hamilton, H., Starner, T. and Hamilton, S. (2005) Development of an American Sign Language Game for Deaf Children. In Proceedings of the Conference on Interaction Design and Children IDC'05
52. Hornof, A. J. and Cavender, A. (2005) EyeDraw: Enabling Children with Severe Motor Impariments to Draw with Their Eyes. In Proceedings of the Conference on Human Factors in Computing Systems CHI'05, ACM, pp. 161 - 170
53. Hourcade, J. P., Bederson, B. B., Druin, A., Rose, A., Farber, A. and Takayama, Y. (2002) The International Children's Digital Library: Viewing Digital Books Online. In Proceedings of the Conference on Interaction Design and Children IDC'02
54. Hourcade, J. P., Bederson, B. B., Druin, A., Rose, A., Farber, A. and Takayama, Y. (2003) The International Children's Digital Library: Viewing Digital Books Online. Interacting with Computers, Vol. 15, pp. 151-167
55. Hourcade, J. P., Bederson, B. B., Druin, A. and Guimbretière, F. (2004) Differences in pointing task performance between preschool children and adults using mice. ACM Transactions on Computer-Human Interaction (TOCHI), Vol. 11(4), pp. 357 - 386
56. Höysniemi, J., Hämäläinen, P. and Turkki, L. (2002) Using Peer Tutoring in Evaluating Usability of Physically Interactive Computer Game with Children. In Proceedings of the Conference on Interaction Design and Children IDC'02
57. Höysniemi, J., Hämäläinen, P. and Turkki, L. (2003) Using Peer Tutoring in Evaluating the Usability of a Physically Interactive Computer Game with Children. Interacting with Computers, Vol. 15, pp. 203-225
58. Höysniemi, J., Hämäläinen, P. and Turkki, L. (2004) Wizard of Oz prototyping of computer vision based action games for children. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 27 - 34
59. Inkpen, K. M. (2001) Drag-and-Drop versus Point-and-Click Mouse Interaction Styles for Children. ACM Transactions on Computer-Human Interaction, Vol. 8 (1), pp. 1-33
60. Isomursu, M., Isomursu, P. and Still, K. (2004) Capturing tacit knowledge from young girls. Interacting with Computers, Elsevier, Volume 16(3), pp. 431 - 449
61. Iversen, O. S. (2002) Designing with Children. The Video Camera as an Instrument of Provocation. In Proceedings of the Interaction Design and Children (IDC'02)
62. Jacko, J. A. (1996) The Identifiability of Auditory Icons for Use in Educational Software for Children. Interacting with Computers, Vol. 8 (2), pp. 121-133
63. Jensen, J. J. and Skov, M. B. (2005) A Review of Research Methods in Children's Technology Design. In Proceedings of the Conference on Interaction Design and Children IDC'05
64. Johnson, M. P., Wilson, A., Blumberg, B., Kline, C. and Bobick, A. (1999) Sympathetic Interfaces: Using a Plush Toy to Direct Synthetic Characters. In Proceedings of the Conference on Human Factors in Computing Systems CHI'99, ACM, pp. 152 - 158
65. Jones, C., McIver, L., Gibson, L. and Gregor, P. (2003) Experiences Obtained from Designing with Children. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 69 - 74
66. Jung, Y., Persson, P. and Blom, J. (2005) DeDe: Design and Evaluation of a Context-Enhanced Mobile Messaging System. In Proceedings of the Conference on Human Factors in Computing Systems CHI'05, ACM, pp. 351 – 360
67. Kaplan, N. and Chisik, Y. (2005) Reading Alone Together: Creating Sociable Digital Library Books. In Proceedings of the Conference on Interaction Design and Children IDC'05
68. Kaplan, N., Chisik, Y., Knudtzon, K., Kulkarni, R., Moulthrop, S., Summers, K. and Weeks, H. (2004) Supporting sociable literacy in the international children's digital library. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 89 - 96
69. van Kesteren, I. E. H., Bekker, M. M., Vermeeren, A. P. O. S. and Lloyd, P. A. (2003) Assessing Usability Evaluation Methods on Their Effectiveness to Elicit Verbal Comments from Children Subjects. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 41 - 49

70. Kim, S-H., Chung, A., Ok, J-H., Myung, I-S., Kang, H. J., Woo, J-K. K. and Kim, M. J., (2004) Communication enhancer—appliances for better communication in a family. Personal and Ubiquitous Computing, Springer-Verlag, Vol. 8(3-4), pp. 221 - 226
71. Kindborg, M. (2002) Comics, Programming, Children, and Narratives. In Proceedings of the Conference on Interaction Design and Children IDC'02
72. Knudtzon, K., Druin, A., Kaplan, N., Summers, K., Chisik, Y., Kulkarni, R., Moulthrop, S., Weeks, H. and Bederson, B. (2003) Starting an Intergenerational Technology Design Teams: A Case Study. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 51 - 58
73. Labrune, J.-P. and Mackay, W. (2005) Tangicam: Exploring observation tools for children. In Proceedings of the Conference on Interaction Design and Children IDC'05
74. Lamberty, K. K. and Kolodner, J. L. (2005) Camera Talk: Making the Camera a Partial Participant. In Proceedings of the Conference on Human Factors in Computing Systems CHI'05, ACM, pp. 839 - 848
75. Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A. and Bhogal, R. S. (1997) The Persona Effect: Affective Impact of Animated Pedagogical Agents. In Proceedings of the Conference on Human Factors in Computing Systems CHI'97, ACM, pp. 359 - 366
76. Lewis, C., Brand, C., Cherry, G. and Rader, C. (1998) Adapting User Interface Design Methods to the Design of Educational Activities. In Proceedings of the Conference on Human Factors in Computing Systems CHI'98, ACM, pp. 619-626
77. Ling, R. (2001) "We Release Them Little by Little": Maturation and Gender Identity as Seen in the Use of Mobile Technology. Personal and Ubiquitous Computing, Vol. 5, pp. 123-136
78. Loh, B., Radinsky, J., Russell, E., Gomez, L. M., Reiser, B. J. and Edelson, D. C. (1998) The Progress Portfolio: Designing Reflective Tools for a Classroom Context. In Proceedings of the Conference on Human Factors in Computing Systems CHI'98, ACM, pp. 627 - 634
79. Louca, L. (2005) The Syntax or the Story Behind it? A Usability Study of Student Work with Computer-Based Programming Environments in Elementary Science. In Proceedings of the Conference on Human Factors in Computing Systems CHI'05, ACM, pp. 849 - 858
80. Lumbreras, M. and Sánchez, J. (1999) Interactive 3D Sound Hyperstories for Blind Children. In Proceedings of the Conference on Human Factors in Computing Systems CHI'99, ACM, pp. 318-325
81. MacFarlane, S., Sim, G. and Horton, M. (2005) Assessing Usability and Fun in Educational Software. In Proceedings of the Conference on Interaction Design and Children IDC'05
82. Markopoulos, P. and Bekker, M. (2002) How to Compare Usability Testing Methods with Children Participants. In Proceedings of the Conference on Interaction Design and Children IDC'02
83. Markopoulos, P. and Bekker, M. (2003) On the Assessment of Usability Testing Methods for Children. Interacting with Computers, Vol. 15, pp. 227-243
84. Marshall, P., Price, S. and Rogers, Y. (2003) Conceptualizing Tangibles to Support Learning. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 101 - 109
85. McElligott, J. and van Leeuwen, L. (2004) Designing sound tools and toys for blind and visually impaired children. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 65 - 72
86. McNerney, T. S. (2004) From turtles to Tangible Programming Bricks: explorations in physical language design. Personal and Ubiquitous Computing, Springer-Verlag, Vol. 8(5), pp. 326 - 337
87. Milne, S., Gibson, L., Gregor, P. and Keighren, K. (2003) Pupil Consultation Online: Developing a Web-Based Questionnaire System. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 127 - 133
88. Moher, T., Johnson, A., Ohlsson, S. and Gillingham, M. (1999) Bridging Strategies for VR-based Learning. In Proceedings of the Conference on Human Factors in Computing Systems CHI'99, ACM, pp. 536 - 543
89. Montemayor, J., Druin, A., Farber, A., Simms, S., Churaman, W. and D'Amour, A. (2002) Physical Programming: Designing Tools for Children to Create Physical Interactive Environments. In Proceedings of the Conference on Human Factors in Computing Systems CHI'02, ACM, pp. 299 - 306
90. Mäkelä, A., Giller, V., Tscheligi, M. and Sefelin, R. (2000) Joking, Storytelling, Artsharing, Expressing Affection: A Field Trial of how Children and Their Social Network Communicate with Digital Images in Leisure Time. In Proceedings of the Conference on Human Factors in Computing Systems CHI'00, ACM, pp. 548 - 555

91. Oosterholt, R., Kusano, M. and de Vries, G. (1996) Interaction Design and Human Factors Support in the Development of a Personal Communicator for Children. . In Proceedings of the Conference on Human Factors in Computing Systems CHI'96, ACM, pp. 450 - 457
92. Ovaska, S., Hietala, P. and Kangassalo, M. (2003) Electronic Whiteboard in Kindergarten: Opportunities and Requirements. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 15 - 22
93. Oviatt, S., Darves, C. and Coulston, R. (2004) Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. ACM Transactions on Computer-Human Interaction (TOCHI), Vol. 11(3), pp. 300 - 328
94. Paiva, A., Andersson, G., Höök, K., Mourão, D., Costa, M. and Martinho, C. (2002) SenToy in FantasyA: Designing an Affective Sympathetic Interface to a Computer Game. Personal and Ubiquitous Computing, Vol. 6, pp. 378-389
95. Parés, N., Carreras, A., Durany, J., Ferrer, J., Freixa, P., Gómez, D., Kruglanski, O., Parés, R., Ignasi Ribas, J., Soler, M. and Sanjurjo, À. (2005) Promotion of Creative Activity in Children with Severe Autism Through Visuals. In Proceedings of the Conference on Interaction Design and Children IDC'05
96. Price, S., Rogers, Y., Scaife, M., Stanton, D. and Neale, H. (2002) Using 'Tangibles' to Promote Novel Forms of Playful Learning. In Proceedings of the Conference on Interaction Design and Children IDC'02
97. Price, S., Rogers, Y., Scaife, M., Stanton, D. and Neale, H. (2003) Using 'Tangibles' to Promote Novel Forms of Playful Learning. Interacting with Computers, Vol. 15, pp. 169-185
98. Rader, C., Brand, C. and Lewis, C. (1997) Degrees of comprehension: children's understanding of a visual programming environment. In Proceedings of the Conference on Human Factors in Computing Systems CHI'97, ACM, pp. 351 - 358
99. Raffle, H. S., Parkes, A. J. and Ishii, H. (2004) Topobo: a constructive assembly system with kinetic memory. In Proceedings of the Conference on Human Factors in Computing Systems (CHI), ACM Press, pp. 647 - 654
100. Randell, C., Price, S., Rogers, Y., Harris, E. and Fitzpatrick, G. (2004) The Ambient Horn: designing a novel audio-based learning experience. Personal and Ubiquitous Computing, Springer-Verlag, Vol. 8(3-4), pp. 177 - 183
101. Read, J. C., MacFarlaine, S. and Casey, C. (2003) What's Going On? Discovering what Children Understand about Handwriting Recognition Interfaces. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 135 - 140
102. Read, J. C., MacFarlane, S. and Gregory, P. (2004) Requirements for the design of a handwriting recognition based writing interface for children. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 81 - 87
103. Resnick, M. (2005) Some Reflections on Designing Construction Kits for Kids. In Proceedings of the Conference on Interaction Design and Children IDC'05
104. Resnick, M., Martin, F., Berg, R., Borovoy, R., Colella, V., Kramer, K. and Silverman, B. (1998) Digital Manipulatives: New Toys to Think With. In Proceedings of the Conference on Human Factors in Computing Systems CHI'98, ACM, pp. 281 - 287
105. Risden, K., Czerwinski, M., Worley, S., Hamilton, L., Kubiniec, J., Hoffman, H., Mickel, N. and Loftus, E. (1998) Interactive Advertising: Patterns of Use and Effectiveness. In Proceedings of the Conference on Human Factors in Computing Systems CHI'98, ACM, pp. 219 - 224
106. Robertson, J. and Good, J. (2003) Ghostwriter: A Narrative Virtual Environment for Children. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 85 - 91
107. Robertson, J. and Good, J. (2004) Children's narrative development through computer game authoring. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 57 - 64
108. Rode, J. A., Stringer, M., Toye, E. F., Simpson, A. R. and Blackwell, A. F. (2003) Curriculum-Focused Design. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 119 - 126
109. Rogers, Y., Price, S., Fitzpatrick, G., Fleck, R., Harris, E., Smith, H., Randell, C., Muller, H., O'Malley, C., Stanton, D., Thompson, M. and Weal, M. (2004) Ambient wood: designing new forms of digital augmentation for learning outdoors. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 3 - 10
110. Ryokai, K., Marti, S., Ishii, H. (2004) I/O brush: drawing with everyday objects as ink. In Proceedings of the Conference on Human Factors in Computing Systems (CHI), ACM Press, pp. 303 - 310

111. Sadler Takach, B. and Varnhagen, C. (2002) Partnering with Children to Develop Design Guidelines for an Interactive Encyclopedia. In Proceedings of the Conference on Interaction Design and Children IDC'02
112. Scaife, M., Rogers, Y., Aldrich, F. and Davies, M. (1997) Designing for or Designing with? Informant Design for Interactive Learning Environments. In Proceedings of the Conference on Human Factors in Computing Systems CHI'97, ACM, pp. 343 - 350
113. Scaife, M. and Rogers, Y. (2001) Informing the Design of a Virtual Environment to Support Learning in Children. International Journal of Human-Computer Studies, Vol. 55, pp. 115-143
114. Sharples, M., Corlett, D. and Westmancott, O. (2002) The Design and Implementation of a Mobile Learning Resource. Personal and Ubiquitous Computing, Vol. 6, pp. 220-234
115. Sheehan, R. (1999) Incremental Control of a Children's Computing Environment. In Proceedings of the IFIP TC13 Interact '99, IOS Press, pp. 504 - 509
116. Sheehan, R. (2003) Children's Perception of Computer Programming as an Aid to Designing Programming Environment. In Proceedings of the Conference on Interaction Design and Children IDC'03, ACM Press, pp. 75 - 83
117. Sluis, R. J. W. , Weevers, I., van Schijndel, C. H. G. J., Kolos-Mazuryk, L., Fitrianie, S. and Martens, J. B. O. S. (2004) Read-It: five-to-seven-year-old children learn to read in a tabletop environment. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 73 - 80
118. Smith, B. K. and Reiser, B. J. (1998) National Geographic Unplugged: Classroom-Centred Design of Interactive Nature Films. In Proceedings of the Conference on Human Factors in Computing Systems CHI'98, ACM, pp. 424-431
119. Stanton, D., Bayon, V., Neale, H., Ghali, A., Benford, S., Cobb, S., Ingram, R., O'Malley, C., Wilson, J. and Pridmore, T. (2001) Classroom Collaboration in the Design Tangible Interfaces for Storytelling. In Proceedings of the Conference on Human Factors in Computing Systems CHI'01, ACM, pp. 482 - 489
120. Stewart, J., Bederson, B. B. and Druin, A. (1999) Single Display Groupware: A Model for Co-Present Collaboration. In Proceedings of the Conference on Human Factors in Computing Systems CHI'99, ACM, pp. 286 - 293
121. Stringer, M., Toye, E. F., Rode, J. A., Blackwell, A. F. (2004) Teaching rhetorical skills with a tangible user interface. In Proceedings of the 4th International Conference on Interaction Design and Children (IDC), ACM Press, pp. 11 - 18
122. Strommen, E. F., Revelle, G. L., Medoff, L. M. and Razavi, S. (1996) Slow and Steady Wins the Race? Three-Year-Old Children and Pointing Device Use. Behaviour and Information Technology, Vol. 15 (1), pp. 57-64
123. Strommen, E. (1998) When the Interface is a Talking Dinosaur: Learning Across Media with ActiMates Barney. In Proceedings of the Conference on Human Factors in Computing Systems CHI'98, ACM, pp. 288 - 295
124. Strommen, E. and Alexander, K. (1999) Emotional Interfaces for Interactive Aardvarks: Designing Affect into Social Interfaces for Children. In Proceedings of the Conference on Human Factors in Computing Systems CHI'99, ACM, pp. 528 - 535
125. Taxén, G, Druin, A., Fast, C. and Kjellin, M. (2001) KidStory: A Technology Design Partnership with Children. Behaviour and Information Technology, Vol. 20 (2), pp. 119-125
126. Vermeeren, A. P. O. S., van Kesteren, I. E. H. and Bekker, M. M. (2003) Managing the Evaluator Effect in User Testing. In Proceedings of the IFIP TC13 Interact '03, IOS Press, pp. 647 - 654
127. Weiss, P. L., Whiteley, C. P., Treviranus, J. and Fels, D. I. (2001) PEBBLES: A Personal Technology for Meeting Educational, Social and Emotional Needs of Hospitalised Children. Personal and Ubiquitous Computing, Vol. 5, pp. 157-168
128. Williams, M., Jones, O. and Fleuriot, C. (2003) Wearable Computing and the Geographies of Urban Childhood - Working with Children to Explore the Potential of new Teechnology. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 111 - 118
129. Williams, M., Jones, O., Fleuriot, C. and Wood, L. (2005) Children and Emerging Wireless Technologies: Investigating the Potential for Spacial Practice. In Proceedings of the Conference on Human Factors in Computing Systems CHI'05, ACM, pp. 819 - 828
130. Wyeth, P. and Wyeth, G. (2001) Electronic Blocks: Tangible Programming Elements for Preschoolers. In Proceedings of the IFIP TC13 Interact '01, IOS Press, pp. 496 – 503
131. Wyeth, P. and Purchase, H. C. (2003) Using Developmental Theories to Inform the Design of Technology for Children. In Proceedings of the Conference on Interaction Design and Children IDC'03, pp. 93 – 100

132. Zuckerman, O., Arida, S. and Resnick, M. (2005) Extending Tangible Interfaces for Education: Digital Montessori-Inspired Manipulatives. In Proceedings of the Conference on Human Factors in Computing Systems CHI'05, ACM, pp. 859 - 868